machine learning & knowledge extraction

Article

Beyond Forecast Accuracy: Evaluating the Error–Profit Paradox in AI-Based Copper Price Prediction

László Vancsura [1], Tibor Tatay [2,*] and Tibor Bareith [3,4]

[1] Department of Marketing and Supply Chain Analysis, Institute of Agricultural and Food Economics, Hungarian University of Agriculture and Life Sciences, 7400 Kaposvár, Hungary; vancsura.laszlo@uni-mate.hu
[2] Department of Statistics, Finances and Controlling, Széchenyi István University, 9026 Győr, Hungary
[3] ELTE, Centre for Economic and Regional Studies, Institute of Economics, 1097 Budapest, Hungary; bareith.tibor@krtk.elte.hu
[4] Economic Geography and Urban Marketing Centre, John von Neumann University, 1117 Budapest, Hungary
* Correspondence: tatay@sze.hu

Abstract

Copper is a strategically important commodity whose price dynamics are increasingly affected by structural changes, geopolitical shocks, and the global energy transition. These conditions create substantial challenges for forecasting models and provide a useful setting for evaluating the practical value of machine learning predictions. This study compares statistical and artificial intelligence-based forecasting models for copper price prediction under different market regimes and structural break conditions. Model performance is assessed using a multi-dimensional evaluation framework that combines statistical accuracy (MAPE), dynamic pattern reproduction (Taylor diagrams and time-lagged cross-correlation analysis), and the economic performance of forecast-driven trading strategies. The results reveal a consistent error–profit paradox: models with the highest statistical forecasting accuracy do not necessarily generate the best trading outcomes. In several cases, models with larger prediction errors achieve superior economic performance because they capture directional market dynamics more effectively. The analyses further show that structural breaks substantially alter model rankings and predictive usefulness, highlighting the importance of regime-aware evaluation. These findings suggest that forecast accuracy alone provides an incomplete assessment of model quality in financial and commodity forecasting applications. The study contributes to machine learning evaluation research by proposing an integrated framework that jointly considers predictive accuracy, temporal dynamics, model robustness, and economic utility, thereby offering a more comprehensive approach to assessing forecasting systems in real-world decision-making environments.

Keywords: copper; time series forecasting; machine learning; error–profit paradox

Check for updates

Academic Editor: Oliver Hinz

Received: 13 June 2026
Revised: 10 July 2026
Accepted: 13 July 2026
Published: 15 July 2026

1. Introduction

Copper has long been regarded as one of the most strategically important industrial commodities due to its critical role in electricity transmission, the electronics industry, and infrastructure development. In recent years, its economic significance has increased further as global economies have accelerated their transition toward electrification, renewable energy deployment, and digital infrastructure expansion [1–3]. Copper is an essential input in the production of electric vehicles, renewable energy systems, energy storage technologies, and the infrastructure underpinning the digital economy, including data centers and telecommunications networks [4,5]. Consequently, developments in the copper market have become increasingly intertwined with the structural transformation of the global economy, particularly through the green transition and the expansion of the digital economy. The rapid growth of electrification technologies and digital infrastructure has substantially increased global demand for copper, creating new challenges for both market participants and policymakers [6,7]. At the same time, copper market volatility has intensified under the influence of geopolitical tensions, supply chain disruptions, macroeconomic shocks, and financial market uncertainty. Events such as the COVID-19 pandemic, geopolitical conflicts, energy market crises, and inflationary shocks have demonstrated that commodity market dynamics can change rapidly, generating structural breaks and regime shifts in price series [8–10]. In this environment, reliable copper price forecasting has become increasingly important for investors, commodity traders, and industrial stakeholders making investment decisions related to the energy transition and digital infrastructure development.

Forecasting commodity prices has long been a central topic in financial economics and time-series analysis. Traditional approaches, such as ARIMA and other econometric models, have been widely employed in the analysis of commodity price series and continue to serve as important benchmark methods due to their interpretability and transparent assumptions. Over the past decade, however, the emergence of machine learning and deep learning techniques has significantly expanded the forecasting toolkit [11–14]. Models such as Random Forest, Support Vector Regression (SVR), boosting algorithms, and neural network–based architectures, including Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Transformer-based models, are capable of capturing nonlinear relationships and complex temporal patterns in time-series data. As a result, they have been increasingly adopted in financial and commodity market forecasting applications [15–17].

Recent reviews suggest that machine learning approaches frequently outperform traditional econometric models in environments characterized by nonlinear adjustment processes, high volatility, and complex interactions among market variables. In particular, ensemble-learning methods and deep-learning architectures have demonstrated strong predictive capabilities in financial and commodity markets by capturing nonlinear dependencies and noisy sequential patterns [18,19]. Nevertheless, empirical evidence remains mixed. While recurrent architectures such as LSTM have generated promising forecasting and trading results in several financial applications, their advantages may diminish under changing market conditions, evolving market efficiency, and the inclusion of realistic trading constraints [20]. Similarly, more recent attention-based architectures, such as the Temporal Fusion Transformer (TFT), provide enhanced interpretability and multi-horizon forecasting capabilities, yet their empirical superiority often depends on data availability, feature design, and the stability of the underlying market environment [21].

The performance of forecasting models is traditionally evaluated using statistical error measures such as Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics provide important information regarding forecast accuracy and remain the most widely used criteria for model comparison. However, given the highly volatile nature of commodity markets, the frequent occurrence of structural breaks, and the practical use of forecasts in decision-making processes, it is increasingly important to examine whether different models can deliver stable and reliable forecasts across varying market regimes [22–24].

This distinction is particularly relevant in commodity markets, where forecasts are frequently used to support hedging decisions, inventory management, procurement planning, and speculative trading activities. In such settings, a forecasting model may possess substantial economic value even if it does not achieve the lowest statistical forecasting error.

Models capable of identifying turning points, directional persistence, or volatility-adjusted trading opportunities may generate superior decision outcomes despite exhibiting comparatively higher MAPE or RMSE values. Conversely, models with excellent statistical accuracy may produce limited economic benefits if forecast errors occur during critical market reversals or fail to generate actionable trading signals. These observations suggest that the relevant question is not simply which forecasting architecture minimizes statistical error, but whether a model captures economically meaningful price dynamics under changing market conditions. This tension between statistical forecasting accuracy and economic usefulness provides the conceptual foundation for the Error–Profit framework examined in this study.

The present study aims to compare the performance of different forecasting models applied to copper price prediction. The empirical analysis focuses on copper prices, which provide an ideal case study for examining commodities of strategic importance to electrification and the digital economy. Traditional econometric, machine learning, and deep learning models are evaluated using a range of statistical performance measures. The analysis considers multiple forecasting horizons, with particular emphasis on six- and twelve-month horizons, representing different decision-making contexts ranging from short-term trading strategies to medium-term industrial planning. In addition, special attention is devoted to investigating how forecasting performance changes in the presence of structural breaks and market regime shifts.

1.1. Research Gap

Although research on commodity price forecasting has expanded considerably in recent years, the literature remains fragmented in three important respects. First, many studies compare forecasting architectures mainly through statistical error measures, which are useful for measuring average fit but do not directly establish whether forecasts can support profitable or risk-adjusted decisions. Second, empirical results are heterogeneous: recurrent networks, convolutional structures, ensemble methods and transformer-based models have each been reported as strong performers in specific settings, but their rankings often change across assets, horizons, feature sets and volatility regimes [25,26]. Third, relatively few commodity studies integrate structural-break awareness, dynamic-pattern diagnostics and trading performance within the same evaluation design. The research gap therefore emerges not from the absence of AI-based forecasting studies, but from the absence of a unified, regime-aware evaluation framework that connects forecast accuracy, temporal structure and economic utility.

Building on this synthesis, the present study treats copper as a demanding test case for forecast evaluation. The market combines long-run demand pressures linked to electrification and digital infrastructure with short-run shocks generated by macroeconomic, geopolitical and supply-chain disturbances. This setting allows us to examine whether model rankings derived from conventional error metrics remain stable when forecasts are evaluated through dynamic similarity measures and forecast-driven trading outcomes. In this way, the study directly responds to the need for a more analytical bridge between the AI forecasting literature and the practical decision problems faced by commodity-market participants.

1.2. Accordingly, This Study Addresses the Following Research Questions

RQ1: How do market regime shifts and structural breaks affect the performance of different forecasting models used for copper price prediction?

RQ2: To what extent are these models capable of reproducing the statistical structure and dynamics of copper price series, as evaluated through Taylor diagrams and Time-Lagged Cross-Correlation (TLCC) analysis?

RQ3: Is the Error–Profit Paradox evident in the copper market, such that models achieving the highest forecasting accuracy do not necessarily deliver the highest risk-adjusted trading performance?

1.3. The Innovation of the Study Can Be Summarized Through Five Interrelated Contributions

- It develops a regime-aware comparison of traditional econometric, machine learning and deep learning models for copper price prediction, rather than evaluating architectures under a single assumed market environment.
- It extends the empirical evidence on AI-based commodity forecasting to a strategically important industrial metal that is central to electrification, renewable-energy investment and the expansion of digital infrastructure.
- It evaluates robustness under structural breaks and market regime shifts, thereby addressing the instability of model rankings that is frequently observed in volatile commodity markets.
- It combines multiple horizons, dynamic-pattern diagnostics and forecast-driven trading simulation, which allows statistical accuracy to be assessed alongside temporal similarity and economic usefulness.
- It provides direct evidence on the Error–Profit Paradox by testing whether the most accurate copper-price forecasts are also those that produce the strongest risk-adjusted trading outcomes.

2. Literature Review

2.1. Artificial Intelligence in Commodity Market Forecasting

Over the past decade, financial and commodity market forecasting has undergone substantial methodological development, largely driven by the rapid adoption of machine learning and deep learning techniques [27–30]. Recent review studies further confirm that machine learning and deep learning methods have become dominant research directions in financial forecasting, while emphasizing that no single forecasting architecture consistently outperforms all alternatives across different market environments. Instead, forecasting performance depends on market characteristics, data quality, forecasting horizon, and the presence of structural changes [18,19].

Traditional econometric models, such as ARIMA and VAR, often struggle to capture the nonlinear patterns, volatility dynamics, and structural changes that characterize financial time series [31–33]. Consequently, research has increasingly shifted toward deep neural architectures capable of modeling complex temporal dependencies. Nevertheless, the empirical literature can be broadly classified into three groups. The first group comprises classical statistical and econometric approaches, including ARIMA, VAR, and volatility models, which remain attractive because of their transparency and interpretability. The second group consists of machine learning algorithms, such as Support Vector Regression (SVR), Random Forest (RF), boosting methods, and other ensemble techniques, which are capable of modelling nonlinear relationships but whose performance often depends on feature engineering, sample characteristics, and the treatment of non-stationarity [15]. The third group includes deep learning architectures specifically designed for sequential data, such as recurrent, convolutional, and transformer-based models.

Among the most widely applied approaches are Recurrent Neural Networks (RNNs) and their advanced variants, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. These architectures are specifically designed to capture

https://doi.org/10.3390/make8070209

long-range temporal dependencies, a critical feature of financial and commodity market data [34–36]. Empirical studies have demonstrated that LSTM-based models frequently achieve substantially lower forecasting errors than conventional statistical methods, particularly under highly volatile market conditions. However, recent review studies also indicate that the superiority of LSTM-type models is not universal and depends on market conditions, forecasting horizon, and data frequency. Attention-enhanced and CNN-LSTM architectures often rank among the best-performing approaches, although no single model consistently dominates across all forecasting tasks [20,21].

GRU architectures emerged as a simplified alternative to LSTMs, employing fewer parameters while preserving the ability to model long-term dependencies [37–39]. GRU models may be particularly advantageous in applications characterized by limited data availability or where computational efficiency is a relevant consideration. Recent studies further suggest that GRU architectures can incorporate economic consistency into the learning process, for example by embedding the negative elasticity relationship between demand and prices [40].

More advanced architectures have also been introduced in financial time-series forecasting. Temporal Convolutional Networks (TCNs), for example, employ convolutional layers to efficiently process long-term temporal patterns while offering a more stable training process than conventional RNN-based approaches [41–44]. Empirical evidence suggests that TCN models may be particularly effective in forecasting nonlinear and noisy financial time series. Recent developments have increasingly focused on transformer-based architectures, among which the Temporal Fusion Transformer (TFT) has attracted considerable attention. These models can simultaneously handle multivariate time series, exogenous variables, and multiple forecasting horizons [45,46]. Several studies report that TFT models often outperform earlier deep learning approaches, particularly when applied to complex financial datasets [47,48]. The adoption of deep learning techniques has therefore significantly enhanced forecasting capabilities in commodity markets. Nevertheless, the literature increasingly emphasizes that statistical accuracy alone is insufficient for assessing the economic value of forecasting models, particularly when forecasts serve as inputs for trading and investment decisions [49]. Recent applications in commodity markets similarly indicate that forecasting models should be evaluated not only according to conventional statistical accuracy measures but also with respect to their ability to support practical decision-making, including trading, hedging, and risk management [50].

Industrial metals, and copper in particular, occupy a central position in the global economy due to their extensive use in manufacturing, energy infrastructure, and electronics production. Over recent decades, demand for raw materials has increased substantially, driven primarily by global economic growth and industrial expansion [51–54]. According to the United States Geological Survey [55], this trend is partly attributable to technological progress, as many modern technologies—including electric vehicles and renewable energy systems—require large quantities of conductive metals. At the same time, metal price volatility represents a significant economic risk, particularly for countries whose economies depend heavily on commodity exports [56]. Because metals constitute an important source of export revenue for many nations, fluctuations in their prices directly affect macroeconomic stability and economic performance [57].

The dynamics of copper markets are shaped not only by fundamental supply and demand factors but also by broader macroeconomic and financial developments. Previous research has shown that global business cycles, exchange-rate movements—especially fluctuations in the U.S. dollar—and geopolitical uncertainty exert significant influence on copper price behavior [58,59]. Given these complex interactions, copper price forecasting plays an important role for industrial firms, investors, and policymakers alike, as reliable forecasts can contribute to improved risk management and more effective decision-making [60,61]. Price volatility directly affects the profitability of mining, smelting, and manufacturing companies, as well as their investment and risk-management strategies [62].

Recent copper-specific studies further demonstrate that tree-based methods, SVR, neural networks, LSTM, and hybrid deep learning architectures can improve forecasting accuracy relative to traditional statistical benchmarks, particularly when nonlinear dynamics are present. Nevertheless, the reported superiority of individual model classes is not consistent across different sample periods and forecasting horizons, suggesting that market regimes and structural changes substantially influence model performance [25,26]. Moreover, most existing studies continue to emphasize point-forecast accuracy, whereas comparatively little attention has been devoted to assessing whether forecasts also provide economically useful trading or hedging signals.

Overall, the literature indicates that forecasting methodologies for industrial metals have evolved considerably, shifting from traditional statistical techniques toward machine learning and deep learning architectures. However, most existing studies continue to focus on models such as ARIMA, Support Vector Machines (SVMs), conventional neural networks, and LSTM-based architectures. By contrast, the application of more recent temporal deep learning models, including Temporal Convolutional Networks (TCNs) and Temporal Fusion Transformers (TFTs), remains relatively limited in the context of copper price forecasting. Furthermore, relatively few studies have systematically compared statistical, machine learning, and modern deep learning approaches within a unified evaluation framework that simultaneously considers statistical accuracy, dynamic pattern reproduction, statistical significance, and economic performance. This suggests that the adoption of modern time-series architectures represents a promising avenue for future research aimed at improving both forecasting accuracy and the economic usefulness of copper market predictions.

2.2. Structural Breaks and Their Importance in Time-Series Modeling

One of the fundamental characteristics of economic and financial time series is the frequent presence of structural breaks. A structural break occurs when the underlying data-generating process changes over time as a result of events such as economic crises, political developments, or technological transformations [63–65]. Ignoring such breaks may bias statistical estimates and substantially reduce forecasting accuracy. The econometric literature offers a variety of approaches for identifying structural breaks. Among the classical methods is the Chow test, which evaluates the stability of regression parameters when the break date is known a priori [66]. However, its applicability is limited because the timing of structural changes is rarely known in advance in real-world economic processes. To address this limitation, the Zivot–Andrews test was developed, allowing the most significant structural break in a time series to be determined endogenously [67]. Unlike conventional unit root tests, this approach explicitly accounts for structural breaks that might otherwise lead to misleading conclusions regarding stationarity. Empirical applications have demonstrated the effectiveness of the Zivot–Andrews test in detecting breaks associated with major economic shocks, including financial crises and exchange-rate fluctuations [68]. Accounting for structural breaks is particularly important in forecasting applications. Models calibrated under one economic regime may experience a substantial deterioration in predictive performance following a regime shift [69,70]. The literature identifies the failure to account for such regime changes as one of the primary sources of forecasting errors. Consequently, recent research has increasingly focused on integrating structural-break detection mechanisms into machine learning frameworks [71].

2.3. Forecasting Accuracy and Trading Performance

The performance of forecasting models is traditionally evaluated using statistical error metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). While these measures facilitate comparisons across competing models, they do not necessarily reflect the economic value of forecasts [72]. In financial and energy-market applications, the ultimate objective of forecasting is often not merely to predict prices accurately but to support profitable trading decisions. From this perspective, the economic value of a forecasting model depends primarily on its ability to facilitate successful trading strategies.

An increasing body of literature highlights that statistical accuracy and trading performance frequently diverge. Empirical studies have shown that models achieving the best RMSE or MAE values do not necessarily generate the most profitable trading outcomes [73]. This discrepancy arises partly because statistical error measures assess the average magnitude of forecast errors, whereas financial decision-making often depends more heavily on correctly predicting the direction of price movements [74,75]. As a result, several studies advocate alternative evaluation frameworks that are more directly linked to trading applications [76,77].

Examples include direction-based long-short strategies, confidence-filtered trading rules, volatility-scaled position sizing, ensemble signal aggregation, and reinforcement-learning frameworks that directly optimize sequential trading decisions, as summarized by Bhuiyan et al. (2025) [50]. Under such approaches, model performance is assessed not only through statistical errors but also through realized economic outcomes, including returns, drawdowns, win rates, and risk-adjusted performance measures.

These findings suggest that the evaluation of financial forecasting models should rely on an integrated framework that simultaneously considers statistical accuracy and economic usefulness [78–80].

Recent empirical evidence further indicates that machine-learning forecasts can generate substantial economic gains when they successfully capture nonlinear interactions and tradable directional information, even if this does not coincide with the lowest conventional forecasting error. This phenomenon has been demonstrated in financial-market applications by Fischer and Krauss [20] and Gu et al. [81], who show that economic performance may diverge from purely statistical forecast accuracy. Consequently, model evaluation should extend beyond traditional error metrics to include indicators that reflect practical decision-making performance.

This consideration is particularly relevant in commodity markets, where price movements are highly volatile and frequently characterized by regime shifts. In such environments, it is essential to assess the extent to which forecasting models can support practical trading decisions. The literature indicates that machine learning and deep learning approaches have substantially advanced the forecasting of financial and commodity prices, particularly in modeling nonlinear patterns and long-term dependencies [82]. Nevertheless, numerous studies emphasize that commodity markets are often characterized by structural breaks and regime shifts that can significantly affect model stability and out-of-sample performance. Under such conditions, patterns learned from historical data may rapidly lose predictive relevance, posing particular challenges for longer forecasting horizons [83].

Relatively few studies simultaneously examine the performance of different machine learning models in environments characterized by structural breaks and investigate the extent to which their forecasts can be effectively translated into trading decisions. This gap is especially evident in industrial metal markets, such as copper, where prices are influenced simultaneously by macroeconomic cycles, financial factors, and geopolitical shocks.

The Error-Profit Paradox provides a useful conceptual framework for explaining this discrepancy, emphasizing that statistical forecast accuracy and economic usefulness are related but not identical objectives. Accordingly, the present study evaluates the forecasting performance of a range of statistical, machine-learning, and deep-learning models in the copper market, with particular emphasis on periods characterized by structural breaks. The objective is not only to compare the statistical accuracy of alternative forecasting approaches but also to examine the extent to which these models can contribute to practical trading and investment decision-making by combining conventional error metrics with Taylor diagrams, time-lagged cross-correlation analysis, and forecast-driven trading simulation.

3. Materials and Methods

3.1. Data

The analysis focuses on two distinct periods of copper price dynamics. The first interval covers the period from 1 June 2014 to 31 December 2022, representing market conditions before and during the COVID-19 pandemic. The second interval spans 1 June 2017 to 31 December 2025, capturing the post-pandemic copper market characterized by heightened volatility and structural changes. The use of these two samples enables a comparison of model performance across different market regimes, with particular emphasis on the evolution of price dynamics over time. The selection of the 2022 and 2025 calendar years as the focal regimes for our comparative evaluation is driven by profound and distinct macroeconomic and geopolitical transformations in the global commodity markets, which represent ideal stress-testing environments for predictive algorithms. The 2022 regime encapsulates an unprecedented energy crisis triggered by the outbreak of the Russia-Ukraine conflict, coupled with a historic monetary policy pivot as the Federal Reserve initiated an aggressive interest rate hike cycle. These factors combined to inject extreme volatility into base metal markets, fundamentally altering established price velocity and correlation structures for copper. Conversely, the 2025 regime represents a structurally distinct phase of market disruption, characterized by severe supply-side shocks—including protracted operational disruptions in major Latin American copper mines (e.g., Morenci Mine—Arizona/USA, Kennecott Copper Project—Utah/USA)—juxtaposed against highly volatile demand signals from the shifting global green energy transition and fluctuating industrial outputs in major consuming economies. Evaluating model performance across these two entire calendar years ensures that the predictive architectures are stress-tested not merely against isolated price spikes, but against comprehensive, structurally distinct macroeconomic cycles encompassing the pre-shock uncertainty, the immediate crisis, and the subsequent market recalibration. Daily copper futures price data were obtained from the Yahoo Finance database. Since the analysis was conducted within a sequential forecasting framework, the time series were used in their natural chronological order. No data imputation or artificial data augmentation techniques were applied. This approach ensures that the forecasting models rely exclusively on historically observed information, consistent with the methodological principles of time-series forecasting. When interpreting the samples, it is important to distinguish between the full dataset and the subsamples used for descriptive and forecasting purposes. Table 1 summarizes the stationarity and structural break tests performed on the complete available time series covering the period from 1 June 2014 to 31 December 2025.

The stationarity properties of copper prices were examined using several unit root tests (Table 1), including the Augmented Dickey–Fuller (ADF), Phillips–Perron (PP), DFGLS, and KPSS tests, as well as the Zivot–Andrews test, which explicitly accounts for structural breaks. The ADF test produced a test statistic of -0.2714 with a p-value of 0.9295, indicating that the null hypothesis of a unit root cannot be rejected. Similar results were obtained from the Phillips–Perron test, which yielded a statistic of −0.4132 and a p-value of 0.9079, likewise suggesting the presence of a non-stationary process. The DFGLS test further confirmed this conclusion, reporting a statistic of −0.0833 and a p-value of 0.6642, which again does not permit rejection of the unit root null hypothesis. In contrast, the KPSS test (null hypothesis assumes stationarity) returned a statistic of 6.7899 with a p-value of 0.0100, leading to rejection of the stationarity hypothesis. Taken together, the results of these tests provide consistent evidence that the copper price series is non-stationary in levels, a finding that is consistent with the typical characteristics of financial and commodity price series. The Zivot–Andrews test, which allows for endogenous structural breaks, produced a test statistic of −3.2805 and a p-value of 0.8047, indicating that the null hypothesis of a unit root cannot be rejected even after accounting for structural changes. Nevertheless, the test identified two potential breakpoints within the sample period. The first structural break occurred on 9 June 2022, while the second significant breakpoint was detected on 31 July 2025. These findings suggest that copper price dynamics are characterized not only by non-stationarity but also by substantial structural changes over time. The presence of structural breaks is particularly relevant for forecasting applications, as changes in market regimes may affect both the predictive performance and stability of forecasting models. Consequently, evaluating model performance across periods before and after these structural shifts is essential for assessing forecasting robustness under changing market conditions.

Table 1. Results of unit root, structural break and stationarity tests for copper (1 June 2014 to 31 December 2025).

Statistics	Value
ADF stat	−0.2714
ADF p-value	0.9295
KPSS stat	6.7899
KPSS p-value	0.0100
PP stat	−0.4132
PP p-value	0.9079
DFGLS stat	−0.0833
DFGLS p-value	0.6642
ZA stat	−3.2805
ZA p-value	0.8047
Structural break 2022 (Zivot-Andrews)	9 June 2022
Structural break 2025 (Zivot-Andrews)	31 July 2025

Source: authors' research.

3.2. Methods

The study evaluates a diverse set of forecasting models representing three main methodological families: classical statistical models, machine learning algorithms, and deep learning architectures. This selection enables a comprehensive comparison of linear, nonlinear, and sequence-based approaches for modelling copper price dynamics under varying market conditions.

Classical statistical models are included as benchmark approaches due to their strong theoretical foundation and widespread use in time-series forecasting. In particular, the autoregressive integrated moving average (ARIMA) model is employed to capture linear dependencies and temporal autocorrelation in copper price movements. Despite its

https://doi.org/10.3390/make8070209

structural simplicity, ARIMA remains a competitive baseline in commodity forecasting literature, often providing robust performance in relatively stable market environments.

3.2.1. Autoregressive Integrated Moving Average (ARIMA)

The Autoregressive Integrated Moving Average (ARIMA) model represents one of the foundational approaches in time-series analysis and has long been applied in commodity and metal price forecasting due to its mathematical rigor and interpretability. The underlying assumption of ARIMA is that the future evolution of a time series (y_t) can be modeled as a linear function of past observations and previous forecast errors, provided that the series is rendered stationary. The framework consists of three components: the autoregressive (AR) component captures persistence in past values; the integrated (I) component removes stochastic trends through differencing; and the moving average (MA) component models the gradual dissipation of random shocks. In industrial metal markets, particularly copper, ARIMA is well suited to capturing inertia effects and mean-reverting behavior arising from inventory dynamics and market expectations. Mathematically, an ARIMA (p, d, q) process can be expressed using the lag operator (L) as:

$$\Phi(L)\left(1 - L\right)^d y_t = c + \theta(L)\varepsilon_t \tag{1}$$

where the autoregressive polynomial is defined as

$$\Phi(L) = 1 - \sum_{i=1}^{p} \Phi_i L^i \tag{2}$$

and the moving-average polynomial as

$$\theta(L) = 1 - \sum_{j=1}^{q} \theta_j L^j. \tag{3}$$

The term $(1 - L)^d$ denotes differencing of order d, which removes stochastic trends, while ε_t represents a white-noise process, $\varepsilon_t \sim WN(0, \sigma^2)$. In this study, ARIMA serves as a benchmark model against which the performance of more advanced machine learning and deep learning approaches can be assessed. Owing to its low computational requirements and transparent parameterization, ARIMA continues to be regarded as a reference model in forecasting and supply chain management applications [84].

Machine learning models are incorporated to capture nonlinear relationships in the data without relying on explicit assumptions about the underlying data-generating process. The study considers AdaBoost, Random Forest (RF) and Support Vector Regression (SVR) as representative ensemble and kernel-based learning methods.

These models are capable of identifying complex interactions between input features; however, they generally treat observations as independent and identically distributed, which limits their ability to explicitly model sequential dependencies in time-series data. As a result, their performance in capturing long-term temporal structures may be constrained compared to sequence-aware architectures.

3.2.2. Adaptive Boosting (AdaBoost)

Adaptive Boosting (AdaBoost) is an ensemble-based machine learning algorithm that combines multiple weak learners to construct a stronger predictive model. It has been widely applied across various forecasting tasks, including financial and commodity market prediction. The algorithm builds the ensemble iteratively, assigning greater weight to observations that were incorrectly predicted by previous learners. This weighting mechanism enables the model to progressively improve its forecasting performance [85].

The final AdaBoost prediction can be expressed as:

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{4}$$

where $F(x)$ denotes the final forecast, T is the number of boosting iterations (i.e., weak learners), $h_t(x)$ represents the prediction generated by the t-th learner, and α_t denotes the weight assigned to that learner based on its predictive performance. The key advantage of AdaBoost lies in its ability to combine information from multiple weak learners while assigning greater importance to more accurate predictors. As a result, the method can effectively capture nonlinear patterns and complex data structures. This characteristic is particularly valuable in financial time-series forecasting, where commodity prices frequently exhibit high volatility and nonlinear behavior. In the present study, AdaBoost is employed to forecast copper prices, allowing its performance to be compared directly with both traditional statistical models and deep learning architectures.

3.2.3. Random Forest Regression (RFR)

Random Forest (RF) is an ensemble learning method that combines multiple decision trees to achieve higher predictive performance than individual models. The algorithm is based on the principle that each decision tree is trained on a randomly selected bootstrap sample of the dataset. During model construction, binary recursive partitioning (BRP) is applied, whereby a random subset of input variables is evaluated at each split to identify the optimal partition. This procedure is recursively repeated within each partition until terminal nodes are reached.

The configuration of a Random Forest model is primarily determined by three parameters: the number of trees (n), the set of candidate predictors (K), and the maximum tree depth (J). Individual trees function as weak learners, denoted by τ_i, while the final prediction is obtained through aggregation of their outputs [86].

For a new observation, the predicted value is given by [87]:

$$\hat{R}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{r}_i(x) \tag{5}$$

where $\hat{r}_i(x)$ represents the prediction generated by the i-th decision tree. By averaging across multiple trees, Random Forest effectively reduces variance and improves generalization performance, making it particularly suitable for noisy commodity-market data characterized by complex nonlinear relationships.

3.2.4. Support Vector Regression (SVR)

Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) framework designed specifically for continuous-value prediction tasks, such as commodity price forecasting in environments characterized by substantial complexity and nonlinear dynamics. The methodology is based on the principle of structural risk minimization, aiming to construct a function that achieves the widest possible margin while maintaining predictive accuracy. In financial applications, SVR is particularly effective at filtering irrelevant noise and enhancing forecast stability. The approach maps input observations $x_i \in \mathbb{R}^d$ into a higher-dimensional feature space H through a transformation function Φ, where nonlinear relationships can be represented as linear structures. The optimal regression function is subsequently determined through quadratic optimization procedures [88].

The decision function can be expressed as:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right) \tag{6}$$

where $K(x, x_i)$ denotes the kernel function, α_i are the support vector coefficients, y_i represents the target value, and b is the intercept term. The flexibility of SVR is largely derived from the choice of kernel functions, such as the radial basis function (RBF), polynomial, or sigmoid kernels, which enable the model to capture nonlinear dependencies in the data. SVR is particularly effective in high-dimensional settings; however, careful selection of kernel specifications and regularization parameters is essential to prevent overfitting and ensure robust out-of-sample performance [89].

Deep learning models are included due to their ability to explicitly model temporal dependencies and nonlinear dynamics in sequential data. The study considers Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Temporal Convolutional Network (TCN), DLinear, and Temporal Fusion Transformer (TFT). Recurrent architectures such as LSTM and GRU incorporate memory mechanisms that allow them to retain and update relevant historical information over time, making them particularly suitable for financial and commodity time series characterized by volatility clustering and structural changes. TCN extends this capability using convolutional filters to capture long-range dependencies, while DLinear introduces a decomposition-based structure that separates trend and seasonal components. TFT further enhances interpretability by combining attention mechanisms with sequence modelling, enabling dynamic feature selection across time.

3.2.5. DLinear

The DLinear model challenges the growing complexity of Transformer-based forecasting approaches by revisiting the effectiveness of linear structures augmented with modern decomposition techniques. The central premise of the model is that sophisticated attention mechanisms may introduce unnecessary noise, whereas properly configured linear layers can more robustly identify underlying trends. DLinear decomposes a time series into trend and seasonal components using a moving-average filter and subsequently processes these components through separate weighted linear transformations [90]:

$$\hat{y} = W_t x_{trend} + W_s x_{season} + b \tag{7}$$

where W_t and W_s denote the weight matrices associated with the trend and seasonal components, respectively. This parsimonious architecture offers a high degree of robustness and interpretability, characteristics that are often particularly valuable in economic and managerial decision-making contexts. Furthermore, its relatively low computational requirements make DLinear well suited for large-scale scenario analyses and sensitivity testing.

3.2.6. Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a streamlined alternative to the Long Short-Term Memory (LSTM) architecture, combining memory and gating mechanisms within a more compact framework. The GRU relies on two principal gates: the update gate (z_t) and the reset gate (r_t), which jointly regulate the information contained in the hidden state (h_t). By reducing the number of trainable parameters, GRU often achieves predictive performance comparable to that of LSTM while requiring lower computational resources and exhibiting faster convergence. These characteristics make it particularly attractive for real-time industrial decision-support systems. In commodity markets, GRU effectively models nonlinear price dynamics while maintaining stable training behavior. Its simpler architecture also reduces the risk of overfitting when only limited data are available [91].

The model is defined by the following equations:

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \tag{8}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \tag{9}$$

$$\widetilde{h}_t = \tanh(W[r_t \odot h_{t-1}, x_t] + b) \tag{10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t \tag{11}$$

From a managerial perspective, GRU offers an attractive balance between model complexity and forecasting accuracy. Its inclusion in this study enables a direct comparison among gated recurrent architectures and facilitates an assessment of whether the additional complexity of LSTM is justified in the context of industrial metal price forecasting.

3.2.7. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) network is an extension of the standard recurrent neural network (RNN) architecture, specifically designed to address long-term temporal dependencies and the vanishing gradient problem [92]. Its defining feature is the memory cell (C_t), which allows information to be retained or selectively discarded through dedicated gating mechanisms.

For copper futures prices, such long-term dependencies may arise from multi-year production cycles, strategic investment decisions, and persistent macroeconomic trends. Information flow within the network is regulated through three gates: the forget gate (f_t), which determines which information should be discarded; the input gate (i_t), which controls the incorporation of new information; and the output gate (o_t), which governs the information transferred to the hidden state.

The LSTM architecture can be represented as follows:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{12}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{13}$$

$$\widetilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{14}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \tag{15}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{16}$$

$$h_t = o_t \odot \tanh(C_t) \tag{17}$$

where σ denotes the sigmoid activation function and $\odot$ represents element-wise multiplication. From an economic perspective, the ability of LSTM networks to capture persistent effects may facilitate more reliable medium-term forecasts, which are particularly valuable for procurement planning and cost management. Although computationally intensive, their robustness and adaptability to volatile commodity-market conditions justify their widespread use. In this study, LSTM is included to evaluate whether the introduction of memory cells yields measurable improvements in forecasting performance for complex commodity price dynamics [93].

https://doi.org/10.3390/make8070209

3.2.8. Temporal Convolutional Network (TCN)

The Temporal Convolutional Network (TCN) represents a modern alternative to recurrent architectures for sequential data modeling. Instead of relying on recurrent connections, TCN employs causal and dilated convolutions to capture temporal dependencies. Its principal advantage over RNN-based models is its ability to perform parallel computation, resulting in more stable gradient propagation and faster training. Through the use of dilated convolutions, TCN can effectively capture long-range temporal dependencies without requiring excessively deep network structures [94]. In the context of copper price forecasting, this architecture can simultaneously model short-term volatility and longer-term cyclical dynamics driven by underlying market fundamentals.

The TCN framework can be formalized as:

$$y_t = x_t + \sum_{l=1}^{L} \sum_{k=0}^{K-1} w_k^{(l)} x_{t-d_l k} \tag{18}$$

where x_t denotes the input, y_t output, L number of convolutional blocks, l layer index, K kernel size, k kernel index, w convolutional weights, and d dilation factor. The network further incorporates residual connections that facilitate efficient training of deeper architectures. From an operational perspective, TCN is particularly well suited to rolling-window forecasting frameworks, while its fixed receptive field contributes to robust performance in the presence of noisy industrial and financial data.

3.2.9. Temporal Fusion Transformer (TFT)

The Temporal Fusion Transformer (TFT) is one of the most comprehensive architectures currently available for time-series forecasting, specifically designed to integrate heterogeneous data sources, including historical observations, known future inputs, and static metadata. TFT combines the sequential modeling capabilities of LSTM networks with the attention mechanisms of Transformer architectures, while employing Gated Residual Networks (GRNs) to dynamically assess the relevance of input variables. This structure enables the model to filter out temporarily irrelevant information and focus on the most influential market drivers [95]:

$$\hat{y}_t = \sum_{\tau=1}^{T} \beta_{t,\tau} \, \mathrm{GRN}\left(\mathrm{LSTM}\left(\sum_{i=1}^{M} \alpha_{t,i} x_{\tau,i}\right)\right) \tag{19}$$

where M denotes the number of input variables, T represents the number of considered time steps, $\beta_{t,\tau}$ are temporal attention weights, and $\alpha_{t,i}$ are variable-selection weights. A key feature of TFT is its multi-horizon attention mechanism, which identifies the historical observations that are most relevant to a particular forecast. This enhances model interpretability and provides a degree of transparency that is often lacking in deep learning applications. Consequently, TFT not only generates forecasts but also offers insights into the underlying drivers of its predictions. Given its ability to model regime shifts, nonlinear dependencies, and complex cyclical patterns, TFT is widely regarded as one of the state-of-the-art approaches for forecasting commodity markets, including copper prices.

The inclusion of models from these three categories ensures that the analysis captures a broad spectrum of forecasting philosophies, ranging from linear statistical assumptions to highly flexible deep learning architectures. This diversity is essential for evaluating whether increasing model complexity leads to consistent improvements in forecasting accuracy and economic performance. Importantly, all models are evaluated under an identical experimental framework, ensuring comparability across methodologies. This design allows the study to isolate differences in predictive performance attributable to model structure rather than data processing or evaluation inconsistencies. The performance of these models is assessed using a multidimensional evaluation framework that integrates statistical accuracy metrics, significance testing, dynamic similarity measures, and economic performance evaluation. This framework is described in the following section.

3.3. Rolling Window, Forecasting Design and Hyperparameters

The hyperparameter configurations employed for each forecasting model are presented in detail in Table 2. Separate optimization procedures were conducted for the 2022 and 2025 market regimes, with each architecture individually calibrated using the complete dataset available for the respective period. During the forecasting process, a 50-day rolling-window approach was implemented. This procedure served a dual purpose: it ensured the continuous incorporation of the most recent market information and functioned as a time-series cross-validation mechanism. Consequently, the risk of overfitting was mitigated while maintaining the temporal stability of the forecasting models. The choice of a 50-day rolling window is motivated by its widespread acceptance as a benchmark for medium-term trend analysis in commodity markets, balancing the need for sufficient training data with the capacity to rapidly adapt to structural breaks. While alternative window lengths could offer additional insights, a comprehensive sensitivity analysis across multiple window specifications lies beyond the scope of this paper due to the high computational complexity of the deep learning architectures; this has been noted as a future research direction. To preserve causality, a strict time-based data partitioning strategy was adopted. For the first sample period, the training dataset covered the interval from 1 June 2014 to 31 December 2021, while the second training period extended from 1 June 2017 to 31 December 2024. The corresponding test sets were restricted to the calendar years 2022 and 2025, respectively. Random sampling procedures were deliberately avoided in order to preserve the temporal structure of the time series. The forecasting horizons of 6 and 12 months were selected to reflect the operational realities of corporate and institutional decision-makers in the copper supply chain. These horizons correspond directly to standard semi-annual and annual corporate budgeting, resource procurement, and financial hedging cycles, where long-term directional accuracy is paramount for risk mitigation. Hyperparameter tuning was performed using grid search, with each validation step relying exclusively on historical observations. This approach ensured the complete elimination of information leakage (look-ahead bias) between training and evaluation phases. For neural network and hybrid architectures, an early stopping strategy was employed based on the evolution of the Mean Absolute Error (MAE) measured on the validation set. Training was terminated once no further improvement in validation performance was observed, and the model weights corresponding to the epoch with the lowest validation loss were retained. This procedure enhanced both the generalization capability of the models and the reliability of the resulting forecasts. To ensure adaptability to changing macroeconomic conditions, hyperparameter optimization was conducted separately for each model category and sample period.

Prior to model estimation, all time series were normalized using Min–Max scaling. This transformation maps observations to the $[0, 1]$ interval while preserving the original distributional characteristics and relative differences within the data. Such preprocessing is essential for ensuring numerical stability and facilitating faster convergence in deep learning models.

All computations were performed in a Python 3.14 environment. Model development and evaluation were implemented using the TensorFlow and Scikit-learn frameworks, leveraging their specialized functionalities for time-series validation and hyperparameter optimization.

https://doi.org/10.3390/make8070209

Table 2. Hyperparameter search space and model configurations for the forecasting models.

Model	Parameters	Value	Best Parameters	
			2022	2025
ARIMA	p, d, q	1, 1, 1	-	-
AdaBoost	Number of Estimators	50; 100; 150; 200; 250; 300	100	250
	Base learner	Decision trees regressor	-	-
	Max depth	1; 2; 3; 4; 5; 6; 7; 8; 9; 10	6	8
	Learning rate	0.001; 0.01; 0.1	0.01	0.001
DLinear	Kernel size (multi-scale temporal kernels)	5, 10, 25, 50	25	10
	Dropout rate	0.1; 0.01; 0.001	0.001	0.001
	Batch size	16, 32, 64	64	16
	Activation	Linear	Linear	Linear
	Optimizer	Adam	Adam	Adam
RNN	Hidden Layers	2	2	2
	Hidden layer neuron count	100, 150	150	150
	Batch size	16, 32, 64	32	32
	Epochs	100	-	-
	Activation	ReLU, Tanh, Sigmoid, Linear	Sigmoid	ReLU
	Dropout rate	0.1; 0.01; 0.001	0.001	0.001
	Optimizer	Adam	Adam	Adam
LSTM	Hidden Layers	2	2	2
	Hidden layer neuron count	100, 150	100	150
	Batch size	16, 32, 64	32	32
	Epochs	100	-	-
	Activation	ReLU, Tanh, Sigmoid, Linear	Tanh	Tanh
	Dropout rate	0.1; 0.01; 0.001	0.001	0.001
	Optimizer	Adam	Adam	Adam
GRU	Hidden Layers	2	2	2
	Hidden layer neuron count	100, 150	150	100
	Batch size	16, 32, 64	32	32
	Epochs	100	-	-
	Activation	ReLU, Tanh, Sigmoid, Linear	Tanh	Tanh
	Dropout rate	0.1; 0.01; 0.001	0.001	0.001
	Optimizer	Adam	Adam	Adam
RFR	Number of Estimators	50; 100; 150; 200; 250; 300	200	100
	Max depth	1; 2; 3; 4; 5; 6; 7; 8; 9; 10	6	5
	Min samples leaf	1; 2; 3; 4; 5; 6; 7; 8; 9; 10	7	6
	Min samples split	1; 2; 3; 4; 5; 6; 7; 8; 9; 10	6	4
SVR	Kernel	linear; poly; sigmoid; rbf	rbf	rbf
	C	0.1; 1; 10; 100	100	100
	Gamma	scale; auto	auto	auto
	Epsilon	0.01; 0.1; 0.5	0.01	0.1

Table 2. *Cont.*

Model	Parameters	Value	Best Parameters	
			2022	2025
TCN	Number of layers	2	2	2
	Number of filters	16, 32, 64	32	64
	Kernel size	3	3	3
	Dilation factors	1, 2, 3, 4	3	4
	Dropout rate	0.1; 0.01; 0.001	0.001	0.001
	Batch size	16, 32, 64	16	16
	Activation	ReLU, Tanh, Sigmoid, Linear	Tanh	Tanh
	Optimizer	Adam	Adam	Adam
TFT	Embedding dimension (d_model)	32, 64	32	32
	Number of LSTM layers	2	2	2
	Number of LSTM units	50, 100, 150	100	50
	Number of attention heads	3, 4	3	3
	Dropout rate	0.1; 0.01; 0.001	0.001	0.001
	Batch size	16, 32, 64	32	16
	Activation	ReLU, Tanh, Sigmoid, Linear	Tanh	Tanh
	Optimizer	Adam	Adam	Adam

Source: authors' research.

To ensure methodological transparency and the highest standards of reproducibility, the hyperparameter optimization and training protocols for the deep learning architectures (specifically the GRU, LSTM, TCN, and TFT models) were structured as follows. First, regarding the computational budget, an exhaustive grid search was executed across the entire predefined hyperparameter space (e.g., examining all combinations of hidden units, dropout rates, and learning rates). No artificial constraints or premature termination budgets were applied, ensuring that each architecture reached its optimal predictive capacity within the specified search domain. Second, to prevent any data leakage, a strict temporal data-splitting protocol was implemented utilizing a structural temporal threshold (split date: 1 January 2022 and 1 January 2025). The model training was performed strictly on the historical data prior to this threshold (training set). The hyperparameter selection was subsequently guided by minimizing the Mean Absolute Error (MAE) on the validation set, ensuring that the chosen configuration reflects strong generalization capabilities on unseen data. Third, regarding network initialization, the training framework relied on the robust, optimized default weight initializers embedded within the TensorFlow/Keras environment. The validation set comprised the last 20% of each training window and early stopping with a patience of 12 epochs was applied. To secure exact numerical reproducibility and mitigate potential stochastic variations stemming from initial weight assignments, a standardized global pseudo-random number generator seed was integrated into the final execution scripts, stabilizing the internal state of the layers and ensuring that the documented performance differentials strictly reflect architectural capabilities rather than initialization artifacts.

3.4. Algorithmic Trading Simulation and Risk Management

The evaluation of the predictive models' practical utility is conducted through a parameter-sensitive trading framework, which is systematically benchmarked against a conventional passive (Buy-and-Hold) investment strategy. The methodology's core logic involves translating model-generated price estimations into actionable market positions, subsequently optimizing the risk-adjusted performance—quantified by the Sharpe Ratio—through a multi-dimensional grid of stop-loss (SL), take-profit (TP), and volatility lookback parameters. The strategy operates on the premise of directional convergence. If the model's price estimation for period t ($y_{pred,t}$) exceeds the current spot price (y_t), the algorithm triggers a long position, anticipating an upward correction. Conversely, if the forecasted value is lower than the realized market price, a short position is initiated. The fundamental signal logic and the subsequent log-return calculation are formalized as follows:

If $y_{pred,t} > y_t \rightarrow$ Long Position.

If $y_{pred,t} < y_t \rightarrow$ Short Position.

Log return is calculated as follows:

$$r_t = ln\left(\frac{P_t}{P_{t-1}}\right) \tag{20}$$

To enhance the strategy's robustness, we implemented a dynamic risk management layer based on historical volatility. Three primary hyper-parameters govern this mechanism:

- SL_MULTIPLIER: defines the stop-loss threshold as a function of prior period volatility.
- TP_MULTIPLIER: determines the take-profit level relative to the volatility.
- VOL_LOOKBACK: sets the temporal window (moving standard deviation) used for volatility estimation.

Volatility and limits are calculated as follows:

$$\sigma_t = \mathrm{std}_{\mathrm{rolling}}(r_t, \mathrm{VOL\_LOOKBACK}) \tag{21}$$

$$\mathrm{SL}_t = \mathrm{SL\_MULTIPLIER} \times \sigma_t \tag{22}$$

$$\mathrm{TP}_t = \mathrm{TP\_MULTIPLIER} \times \sigma_t \tag{23}$$

Realized returns are truncated if the price reaches these boundaries. For instance, in a long position, gains exceeding the TP are capped at the TP value, while losses exceeding the SL are limited to $-$SL.

Beyond the baseline logic, two advanced filtering techniques were integrated to assess reliability:

- Confidence-Based Filtering: The strategy calculates a 5-day rolling Mean Absolute Percentage Error (MAPE). A position is only opened if this error metric remains below a 3% threshold, effectively using recent model accuracy as a gatekeeper. This threshold was not chosen arbitrarily but was calibrated to the stylized volatility characteristics of COMEX/LME copper futures, where daily price fluctuations of this magnitude typically fall within the range of ordinary intraday noise and bid-ask spreads. The filter therefore acts as a structural buffer, ensuring that the trading engine opens positions only when a model signals a macroscopically meaningful trend continuation or reversal rather than reacting to statistically insignificant micro-fluctuations.
- Dynamic Sizing: The rolling MAPE serves as a scaling factor for capital allocation. Higher recent error leads to a reduced position size, whereas high historical accuracy justifies a larger exposure.

The program performs static and cumulative performance analysis, calculating the most important statistical indicators for each parameter combination and instrument examined. These include the average daily return (μ), the standard deviation of the daily return (σ), and the Sharpe ratio, which is calculated using the standard annualized formula:

$$\frac{\mu}{\sigma} \times \sqrt{252} \tag{24}$$

In addition, the function calculates the cumulative return, the maximum drawdown (which is the largest difference between the local maxima and minima of the cumulative return curve), and the win rate. The calculations are performed separately for the buy-and-hold benchmark strategy and the dynamically managed strategy, allowing for a multidimensional comparison of the two approaches. During the optimization process, the algorithm iterates through a predefined parameter grid. This grid contains the following values: SL $\in$ {0.5, 1.0, 1.5}, TP $\in$ {1.5, 2.0, 3.0}, and VOL_LOOKBACK $\in$ {10, 20, 30}. The strategy is run for all possible combinations—27 in total—and then the system selects the parameter configuration with the highest Sharpe ratio. The setting determined in this way is stored as a parameter set optimized for the given instrument.

The simulation provides a comparative analysis of cumulative gross returns. In alignment with standard academic practices in initial model validation, the results do not account for fiscal obligations (taxes), brokerage commissions, or slippage. This approach ensures that the performance delta between the Buy-and-Hold benchmark and the active strategy is purely a reflection of the predictive model's information value and the efficiency of the risk management framework. The final output is visualized via cumulative return charts, enabling a multidimensional assessment of whether the predictive models can consistently outperform passive market exposure.

3.5. Performance Evaluation Framework

3.5.1. Forecast Accuracy Metrics

To ensure a comprehensive assessment of forecasting performance, this study employs a multidimensional evaluation framework that integrates four complementary perspectives: (i) forecast accuracy, (ii) statistical significance, (iii) dynamic similarity, and (iv) economic performance. This structure allows for a systematic comparison of forecasting models beyond single-metric evaluation. All models are evaluated under a consistent rolling-window forecasting design, ensuring that performance comparisons are based on identical information sets and out-of-sample predictions.

The most commonly used metrics in the literature for evaluating the predictive models and assessing their accuracy are root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) [96,97].

Root mean squared error (RMSE): this performance indicator shows an estimation of the residuals between actual and predicted values.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{25}$$

where $\hat{y}_i$ is the estimated value produced by the model, y_i is the actual value and n is the number of observations.

Mean Absolute Error (MAE): this indicator measures the average magnitude of the error in a set of predictions.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{26}$$

Mean Absolute Percentage Error (MAPE): this indicator measures the average magnitude of the error in a set of predictions and shows the deviations in percentage.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{27}$$

The reliability and precision of the forecasts increase as the values of the evaluation metrics decrease. From a methodological perspective, RMSE penalizes large forecast errors more heavily because the residuals are squared, making it particularly sensitive to outliers compared with MAE. While both RMSE and MAE measure forecast errors in the original units of the copper futures prices, MAPE expresses the error as a percentage of the observed price, thereby facilitating comparisons across different forecasting models and market regimes. Given that this study evaluates forecasting performance for two distinct periods of the copper futures market characterized by different volatility conditions, MAPE serves as the primary comparative metric because it provides a scale-independent measure of forecast accuracy and enables a consistent assessment of model performance across both sample periods.

3.5.2. Statistical Significance Tests

To assess whether differences in forecasting performance are statistically meaningful, the Model Confidence Set (MCS) procedure and the Diebold–Mariano (DM) test are applied. The MCS procedure identifies a subset of statistically superior models based on iterative elimination of inferior performers. The DM test is used to evaluate pairwise differences in predictive accuracy between competing models. These statistical tests complement error-based metrics by distinguishing meaningful performance differences from random variation.

3.5.3. Dynamic Similarity Analysis

Beyond point-based accuracy, forecasting performance is further evaluated using Taylor diagrams and time-lagged cross-correlation (TLCC) analysis. Taylor diagrams assess the ability of models to reproduce the statistical properties of the observed series, including correlation, standard deviation, and centred root mean squared error. TLCC analysis evaluates the temporal alignment between predicted and observed price series, focusing on how well models capture the timing of market movements across different lag structures. Together, these methods provide insight into the dynamic behaviour of forecasting models beyond static error measures.

3.5.4. Economic Evaluation (Trading Simulation)

To assess the practical relevance of forecasting models, a rule-based trading simulation is implemented. This simulation evaluates whether improvements in statistical and dynamic forecasting performance translate into economic gains. Trading decisions are generated based on forecasted price movements, and performance is evaluated using cumulative returns under a consistent decision rule applied across all models. This stage enables the identification of the error–profit paradox, where lower forecasting error does not necessarily correspond to higher trading profitability.

3.5.5. Cross-Layer Correlation Test

To formally examine the relationship between the accuracy and economic evaluation layers, this study computes Spearman's rank correlation coefficient (ρ) between the MAPE ranking and the Sharpe ratio ranking of the nine forecasting models, separately for each period and forecasting horizon. Trading performance is assessed using two complementary indicators, the Sharpe ratio and the cumulative return, in order to verify that the relationship is not an artefact of a single performance measure. The coefficient is calculated as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{28}$$

where d_i = R(MAPE$_i$) − R(Sharpe$_i$ or Return$_i$) denotes the difference between the accuracy rank and the trading-performance rank of model i, and n is the number of forecasting models compared (n = 9). A positive and statistically significant ρ indicates that less accurate models achieved higher trading performance (consistent with the Error–Profit Paradox), while a negative ρ indicates that forecast accuracy and trading performance are aligned. This test complements the descriptive comparison presented in Sections 4.1 and 4.4 by quantifying the strength and statistical significance of the paradox across market regimes.

Figure 1 summarizes the evaluation logic developed throughout Section 3.5. Forecast accuracy metrics and statistical significance tests (Sections 3.5.1 and 3.5.2) form the numerical and statistical baseline. Dynamic similarity measures (Section 3.5.3) assess the extent to which model forecasts reproduce the structural and temporal properties of the observed price series. The trading simulation (Section 3.5.4) translates forecasting performance into economic outcomes. Finally, the cross-layer correlation test (Section 3.5.5) formally links the accuracy layer to the economic layer, closing the evaluation framework introduced in this section. The empirical results obtained by applying this framework are reported in Section 4.5.

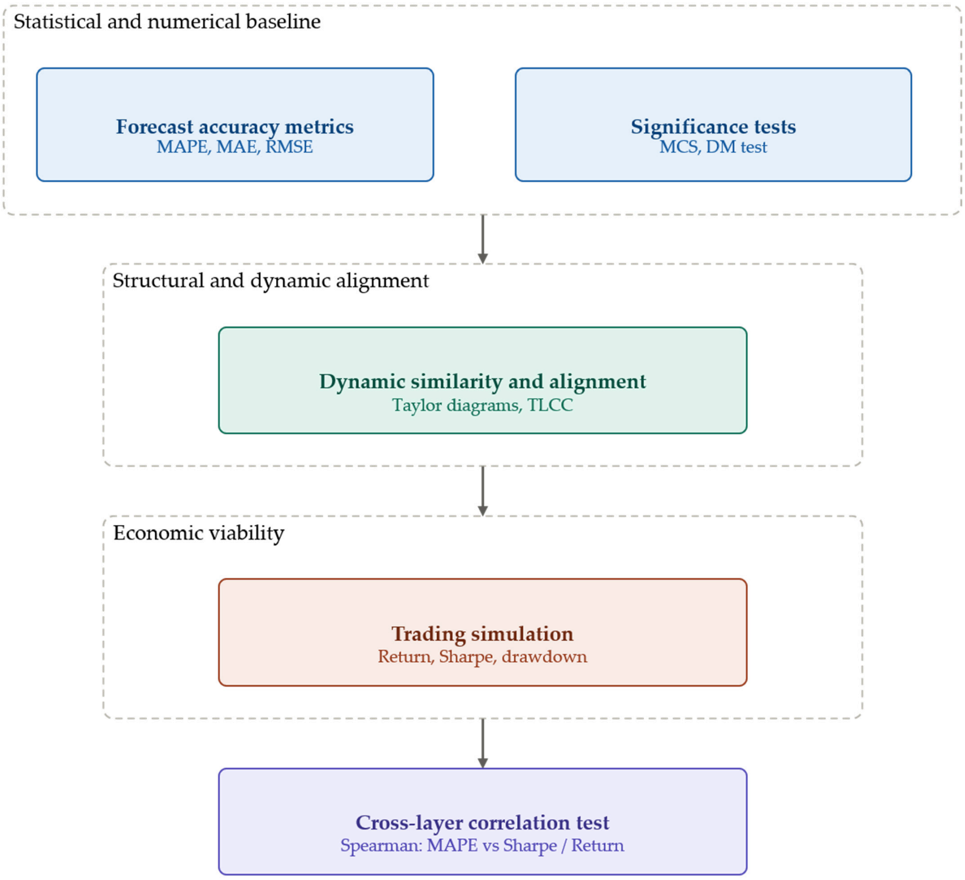

Figure 1. Summary of the multi-dimensional model evaluation framework applied in this study, integrating the four analytical layers. Source: authors' research.

4. Results

The descriptive statistics presented in Table 3 indicate notable differences in both the level and distributional characteristics of copper prices across the two study periods. The average copper price increased from 3.0299 during the first period (2014–2022) to 3.6467 in the second period (2017–2025), suggesting a substantial upward shift in the overall price level. Median values exhibit a similar pattern (2.8385 and 3.7060, respectively), further confirming the movement toward higher price levels over time. The increase in the standard deviation from 0.7041 to 0.7952 suggests a moderate rise in price volatility during the second period. Minimum and maximum values also shifted upward, indicating that the entire distribution moved to a higher price range. A comparison of the quartiles reveals that the central 50% of observations was concentrated at higher price levels in the later period, reinforcing evidence of a broad-based upward adjustment in copper prices. The distribution of prices also changed in terms of shape. During the first period, copper prices exhibited a more pronounced positive skewness (0.8791), indicating a longer right tail and a greater occurrence of unusually high price observations. In contrast, the skewness coefficient in the second period (0.1653) was substantially closer to zero, suggesting a more symmetric distribution. Kurtosis values were negative in both periods and particularly pronounced during the second interval (−1.0294), indicating a flatter, or platykurtic, distribution relative to the normal distribution. This result implies a lower concentration of observations around the mean and thinner tails compared with a Gaussian benchmark. Overall, the descriptive statistics demonstrate that copper prices in the later period were characterized by higher average price levels, moderately increased volatility, and a more balanced distributional structure. These findings are consistent with the substantial changes observed in global commodity markets during the post-pandemic period, which was marked by strong demand growth associated with electrification, energy transition initiatives, and evolving industrial supply chains.

Table 3. Descriptive statistics for Period 1 (1 June 2014 to 31 December 2022) and Period 2 (1 June 2017 to 31 December 2025).

Statistic	Period 1 (2014–2022)	Period 2 (2017–2025)
N	2160	2160
Mean	3.0299	3.6467
Median	2.8385	3.7060
Standard Deviation	0.7041	0.7952
Min	1.9395	2.1195
Max	4.9290	5.7950
25th Percentile (Q1)	2.5918	2.9035
75th Percentile (Q3)	3.2541	4.2991
Skewness	0.8791	0.1653
Kurtosis	−0.1118	−1.0294

Source: authors' research.

4.1. Results of Copper Price Forecasts for 2022 and 2025

Based on the forecasting results for 2022 (Table 4), model performance can primarily be evaluated using the Mean Absolute Percentage Error (MAPE), which expresses relative forecasting error in percentage terms. For the 6-month forecasting horizon, the GRU model achieved the best performance (MAPE = 1.2955%), closely followed by ARIMA (MAPE = 1.3043%). The LSTM and DLinear models also produced low MAPE values between 1.35% and 1.41%, indicating stable and relatively accurate short-term forecasting performance. In contrast, among the classical machine learning models, AdaBoost delivered substantially weaker results (MAPE = 3.1926%), while the Random Forest model also exhibited a relatively high forecasting error (MAPE = 1.8915%). For the 12-month forecasting horizon, a slight deterioration in forecasting accuracy can generally be observed, which is consistent with the increasing uncertainty associated with longer-term predictions. In this case, the GRU model again proved to be the most accurate (MAPE = 1.3524%), while ARIMA and LSTM exhibited similar error levels of approximately 1.40%. Random Forest and AdaBoost performed particularly poorly over the longer horizon, with MAPE values exceeding 3–4%.

Table 4. Performance evaluation indicators and MCS statistic for Period 1 January 2022 to 31 December 2022 (6- and 12-month forecasting horizon).

	6-Month Horizon				12-Month Horizon			
Model	**MAE**	**RMSE**	**MAPE**	**MCS *p*-Value**	**MAE**	**RMSE**	**MAPE**	**MCS *p*-Value**
ARIMA	0.0577	0.0736	1.3043	1.0000	0.0555	0.0716	1.3988	0.5820
AdaBoost	0.1423	0.1845	3.1926	0.0080	0.1451	0.1892	3.6708	0.0010
Dlinear	0.0626	0.0794	1.4148	0.3140	0.0590	0.0752	1.4848	0.0960
GRU	0.0574	0.0743	1.2955	0.7980	0.0538	0.0704	1.3524	1.0000
LSTM	0.0596	0.0755	1.3560	0.7980	0.0559	0.0724	1.4127	0.5820
RFR	0.0818	0.1116	1.8915	0.3140	0.1569	0.2035	4.2447	0.0010
SVR	0.0735	0.0909	1.6688	0.2820	0.0719	0.0919	1.8353	0.0960
TCN	0.0634	0.0832	1.4246	0.3140	0.0615	0.0799	1.5464	0.0960
TFT	0.0748	0.0944	1.6693	0.3140	0.0712	0.0901	1.7858	0.0530

Source: authors' research.

The results of the Model Confidence Set (MCS) procedure provide further insights into the comparison of forecasting models. For the 6-month horizon, the ARIMA model's p-value of 1.0000 indicates that it belongs to the set of best-performing models according to the statistical test. Similarly, GRU and LSTM achieved high p-values (0.7980), suggesting that their performance does not differ significantly from that of the best model. The moderate p-values obtained by DLinear, TCN, SVR, and TFT indicate that these methods remain competitive, although they perform somewhat worse than the top models. In contrast, the very low p-value of AdaBoost (0.0080) suggests statistically inferior performance. For the 12-month horizon, GRU emerged as the leading model within the best-model set (MCS p-value = 1.0000), while ARIMA and LSTM remained among the competitive forecasting methods. Random Forest and AdaBoost again produced very low p-values, confirming that they are unable to provide competitive forecasting accuracy even over longer prediction horizons.

Based on the forecasting results for 2025 (Table 5), the LSTM model achieved the lowest forecasting error for the 6-month horizon (MAPE = 1.4321%), closely followed by GRU (MAPE = 1.4173%) and ARIMA (MAPE = 1.4356%). DLinear exhibited a somewhat higher error (MAPE = 1.5189%), while forecasting errors for TCN and TFT ranged between 1.65% and 1.72%. Among the traditional machine learning approaches, SVR demonstrated moderately weaker performance (MAPE = 2.1777%), whereas AdaBoost and Random Forest showed substantially larger relative errors exceeding 3.6–3.8%. These findings suggest that these methods are less suitable for short-term copper price forecasting. For the 12-

month forecasting horizon, forecasting accuracy generally deteriorated slightly, reflecting the greater uncertainty associated with longer forecasting periods. Neural network-based algorithms continued to perform best. The GRU model achieved a MAPE of 1.4173%, while LSTM and ARIMA exhibited errors around 1.43–1.44%. DLinear produced a forecasting error of 1.5189%. In contrast, the forecasting errors of TCN and TFT increased further. The performance of traditional machine learning methods deteriorated considerably over longer horizons, particularly in the case of AdaBoost and Random Forest, where MAPE values remained above 3%. The Model Confidence Set (MCS) results reinforce the conclusions drawn from the MAPE analysis. For the 6-month horizon, the LSTM model achieved a p-value of 1.0000, indicating that it constitutes the core member of the best-model set according to the statistical test. GRU and ARIMA also achieved relatively high p-values (0.5790 and 0.5650, respectively), indicating that their performance does not differ significantly from that of the best-performing model. The moderate p-values of DLinear, TCN, and SVR suggest that these methods can still be classified as competitive, although their accuracy is somewhat lower. Conversely, the relatively low p-value of AdaBoost (0.0470) indicates statistically inferior performance compared with the best models. For the 12-month horizon, the MCS results reveal a similar pattern. LSTM again achieved the highest p-value (1.0000), while GRU and ARIMA also obtained high p-values (0.8380), indicating stable and competitive forecasting performance. The moderate p-values of DLinear, TCN, and TFT suggest that these methods remain within the set of best-performing models, although their accuracy lags somewhat behind the leading approaches. In contrast, AdaBoost, Random Forest, and SVR delivered significantly weaker forecasting performance over the examined forecasting horizon.

Table 5. Performance evaluation indicators and MCS statistic for Period 2 January 2025 to 31 December 2025 (6- and 12-month forecasting horizon).

	6-Month Horizon				12-Month Horizon			
Model	**MAE**	**RMSE**	**MAPE**	**MCS p-Value**	**MAE**	**RMSE**	**MAPE**	**MCS p-Value**
ARIMA	0.0665	0.1226	1.4356	0.5650	0.0692	0.1226	1.4356	0.8380
AdaBoost	0.1075	0.2951	3.8104	0.0470	0.1965	0.2951	3.8104	0.0190
Dlinear	0.0660	0.1252	1.5189	0.3760	0.0730	0.1252	1.5189	0.4770
GRU	0.0605	0.1224	1.4173	0.5790	0.0684	0.1224	1.4173	0.8380
LSTM	0.0599	0.1212	1.4321	1.0000	0.0692	0.1212	1.4321	1.0000
RFR	0.0922	0.2890	3.6114	0.3760	0.1869	0.2890	3.6114	0.0300
SVR	0.0779	0.1522	2.1777	0.3760	0.1061	0.1522	2.1777	0.0480
TCN	0.0705	0.1328	1.6576	0.3760	0.0793	0.1328	1.6576	0.4770
TFT	0.0684	0.1347	1.7159	0.5230	0.0848	0.1347	1.7159	0.4770

Source: authors' research.

4.2. Statistical Comparison of Model Performance

The results of the Diebold–Mariano (DM) test (Table 6) for the 6-month forecasting horizon in 2022 reveal several significant differences in forecasting performance among the examined models. ARIMA demonstrates significantly better performance than AdaBoost, Random Forest (RFR), SVR, TCN, and TFT, as indicated by positive and statistically significant DM statistics. However, no significant differences are observed between ARIMA and either GRU or LSTM, suggesting that these models provide comparable forecasting accuracy. AdaBoost exhibits significantly weaker performance compared with almost all other methods, as evidenced by the large absolute values and statistical significance of the DM statistics. Among the deep learning algorithms, GRU achieves significantly better performance in several comparisons, particularly against DLinear, RFR, SVR, TCN, and TFT. LSTM also performs favorably; however, compared with GRU and DLinear, statistically significant differences are not observed in all cases. Random Forest proves significantly weaker in many comparisons, especially when evaluated against neural network-based models.

Table 6. Results of DM tests for Period 1 (2022)—6-month horizon.

	ARIMA	AdaBoost	Dlinear	GRU	LSTM	RFR	SVR	TCN	TFT
ARIMA	-								
AdaBoost	7.20 ***	-							
Dlinear	2.57 **	−6.89 ***	-						
GRU	0.26	−7.30 ***	−2.02 **	-					
LSTM	0.56	−7.28 ***	−1.53	0.53	-				
RFR	3.43 ***	−6.37 ***	2.88 ***	3.55 ***	3.59 ***	-			
SVR	3.62 ***	−7.13 ***	2.10 **	4.37 ***	4.20 ***	−2.44 **	-		
TCN	2.20 **	−6.98 ***	0.94	2.67 ***	1.87 *	−2.59 ***	−1.28	-	
TFT	3.20 ***	−7.29 ***	2.27 **	3.56 ***	3.22 ***	−1.42	0.95	2.31 **	-

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Source: authors' research.

For the 12-month forecasting horizon in 2022, the DM test results (Table 7) similarly reveal substantial differences in forecasting performance. ARIMA again demonstrates significantly better performance, particularly compared with AdaBoost, Random Forest (RFR), SVR, TCN, and TFT, as reflected by positive and highly significant DM statistics. However, no statistically significant differences are observed between ARIMA and either GRU or LSTM, suggesting comparable forecasting accuracy even over the longer forecasting horizon. AdaBoost once again performs significantly worse in nearly all pairwise comparisons, as indicated by the strongly negative and highly significant test statistics. GRU significantly outperforms several competing methods, including DLinear, SVR, TCN, and TFT. LSTM remains highly competitive; however, statistically significant differences relative to GRU and ARIMA are not consistently observed. Random Forest demonstrates particularly poor performance compared with most competing methods, as reflected by large and statistically significant DM statistics.

Table 7. Results of DM tests for Period 1 (2022)—12-month horizon.

	ARIMA	AdaBoost	Dlinear	GRU	LSTM	RFR	SVR	TCN	TFT
ARIMA	-								
AdaBoost	9.11 ***	-							
Dlinear	2.11 **	−8.89 ***	-						
GRU	−0.78	−9.22 ***	−3.24 ***	-					
LSTM	0.43	−9.21 ***	−1.68 *	1.86 *	-				
RFR	9.94 ***	1.55	9.83 ***	10.08 ***	10.09 ***	-			
SVR	4.73 ***	−8.76 ***	3.99 ***	5.66 ***	5.49 ***	−9.70 ***	-		
TCN	3.10 ***	−8.76 ***	1.81 *	4.25 ***	2.67 ***	−9.55 ***	−2.83 ***	-	
TFT	4.37 ***	−8.88 ***	3.69 ***	5.40 ***	4.82 ***	−9.29 ***	−0.42	2.86 ***	-

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Source: authors' research.

For the 6-month forecasting horizon in 2025, the DM test results are reported in Table 8. Several statistically significant differences can be observed. ARIMA significantly outperforms AdaBoost, Random Forest (RFR), and SVR, as indicated by positive and significant DM statistics. However, no statistically significant differences are observed between ARIMA and DLinear, GRU, LSTM, TCN, or TFT, suggesting that these models achieve similar forecasting accuracy over the short-term horizon. AdaBoost again exhibits consistently weak performance, performing significantly worse than nearly all competing algorithms. Among the deep learning methods, GRU and LSTM demonstrate significantly better performance in several comparisons, particularly against DLinear, Random Forest, and SVR. Random Forest proves significantly weaker in many cases, especially when compared with neural network-based models. SVR also delivers relatively modest performance, showing significantly poorer results than several competing approaches. For TCN and TFT, most pairwise comparisons reveal no statistically significant differences relative to the leading models, suggesting that their performance remains close to that of the best-performing approaches.

Table 8. Results of DM tests for Period 2 (2025)—6-month horizon.

	ARIMA	AdaBoost	Dlinear	GRU	LSTM	RFR	SVR	TCN	TFT
ARIMA	-								
AdaBoost	4.55 ***	-							
Dlinear	0.69	−4.16 ***	-						
GRU	−0.84	−4.73 ***	−2.61 ***	-					
LSTM	−0.87	−4.73 ***	−2.54 **	−0.56	-				
RFR	3.33 ***	−3.86 ***	2.98 ***	3.53 ***	3.55 ***	-			
SVR	2.21 **	−3.52 ***	1.70 *	2.60 ***	2.56 **	−1.97 **	-		
TCN	1.44	−3.82 ***	1.44	2.79 ***	2.75 ***	−2.47 **	−0.89	-	
TFT	0.35	−5.07 ***	−0.31	1.34	1.39	−3.55 ***	−2.25 **	−2.31 **	-

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Source: authors' research.

For the 12-month forecasting horizon in 2025, the DM test results (Table 9) indicate that ARIMA significantly outperforms AdaBoost, Random Forest (RFR), SVR, and TCN, as demonstrated by positive and statistically significant DM statistics. However, no statistically significant differences can be detected between ARIMA and DLinear, GRU, LSTM, or TFT, suggesting comparable forecasting performance over the longer forecasting horizon. AdaBoost again exhibits consistently weak performance, performing significantly worse than virtually all competing models. Among the deep learning methods, GRU and LSTM significantly outperform several competing approaches, particularly Random Forest, SVR, and TCN. DLinear remains moderately competitive; however, it proves significantly weaker than GRU and LSTM in several comparisons. Random Forest is again among the weakest-performing methods, exhibiting significantly poorer results across numerous pairwise comparisons. Similarly, SVR underperforms relative to the leading models, showing significantly lower forecasting accuracy in several cases. The results for TCN and TFT are mixed. While statistically significant differences are observed relative to some models, in other comparisons their performance does not differ significantly from that of the leading approaches.

Table 9. Results of DM tests for Period 2 (2025)—12-month horizon.

	ARIMA	AdaBoost	Dlinear	GRU	LSTM	RFR	SVR	TCN	TFT
ARIMA	-								
AdaBoost	5.90 ***	-							
Dlinear	0.99	−5.94 ***	-						
GRU	−0.11	−5.98 ***	−1.82 *	-					
LSTM	−0.47	−6.10 ***	−2.38 **	−0.71	-				
RFR	5.68 ***	−7.69 ***	5.71 ***	5.76 ***	5.87 ***	-			
SVR	3.30 ***	−5.86 ***	3.49 ***	3.59 ***	4.27 ***	−5.60 ***	-		
TCN	2.80 ***	−5.76 ***	2.44 **	3.24 ***	3.71 ***	−5.53 ***	−2.56 **	-	
TFT	1.57	−6.37 ***	1.47	1.77 *	2.39 **	−6.12 ***	−4.08 ***	0.29	-

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Source: authors' research.

The empirical results derived from the Diebold–Mariano (DM) and Model Confidence Set (MCS) tests consistently demonstrate the systematic outperformance of recurrent neural architectures (specifically the GRU and LSTM models) relative to ensemble tree-based methods (Random Forest and AdaBoost) across multiple forecasting horizons and distinct market regimes. This persistent, statistically significant performance gap cannot be attributed to numerical coincidence; rather, it is deeply rooted in the underlying mathematical formulations of these competing modeling paradigms. Ensemble tree-based algorithms, despite their capacity to handle high-dimensional feature spaces, operate by partitioning the input space into orthogonal hyperplanes. Consequently, they function as step-wise approximations that inherently lack the structural capacity to extrapolate beyond the exact domain boundaries observed during the training phase. When subjected to severe structural changes—such as the copper market shocks identified by the Zivot-Andrews tests—these algorithms struggle to adapt to out-of-bounds price ranges and shifting volatility regimes, leading to a higher frequency of statistical rejections in the DM pairs. Conversely, recurrent neural networks (RNNs), via their internal hidden states and specialized gating mechanisms (the update and reset gates in GRU, and the forget, input, and output gates in LSTM), are explicitly designed to capture complex, non-linear temporal dependencies. These architectures form a continuous, dynamic memory structure capable of modeling long-memory behavior in time series. This mathematical design allows them to selectively retain historical context from stable periods while rapidly updating their internal weights in response to sudden market regime shifts. This theoretical advantage explains why GRU and LSTM maintain stable, dominant positions within the MCS boundaries even immediately following major structural break points, as they effectively isolate the underlying structural signals from the extreme noise inherent in commodity pricing cycles.

4.3. Structural Break Analysis

4.3.1. Structural Break Analysis for Period 1 (2022)

The heatmap of rolling MAPE values enables the examination of the temporal evolution of forecasting errors for individual models during 2022. As shown in Figure 2, forecasting errors remain relatively low for most models during the first half of the year, as indicated by the predominance of green shades. However, around the structural break marked by the red dashed line, several methods exhibit increasing forecasting errors, suggesting that changes in the market environment affected predictive performance. This effect is particularly pronounced for AdaBoost and Random Forest, where substantial error spikes emerge after the break point, as reflected by the appearance of yellow and red colors. In contrast, deep learning models, especially GRU and LSTM, maintain relatively stable performance, with errors remaining within a lower range throughout most of the period. ARIMA also exhibits a balanced error structure, although minor fluctuations can be observed during the second half of the year. Moderate and occasionally increasing errors are observed for SVR, TCN, and TFT; however, these changes are less pronounced than those of their ensemble-based counterparts.

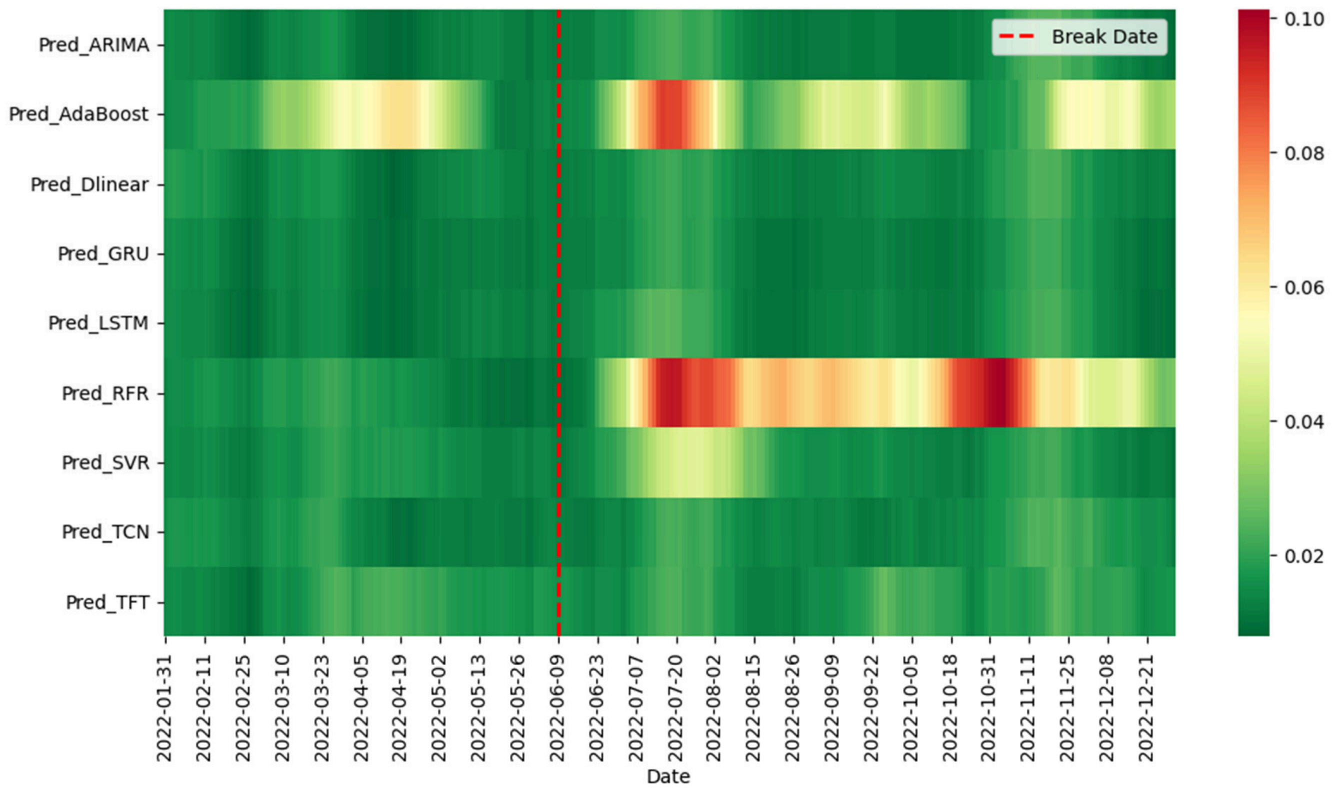

Figure 2. Model-level errors over time (2022) with structural break (5-day rolling MAPE). Source: authors' research.

The autocorrelation function (ACF) plots (Appendix A.1) examine the temporal dependence of forecasting errors before and after the structural break using the 2022 data. For ARIMA, most autocorrelation coefficients remain within the confidence intervals in both periods, indicating that the residuals are close to white noise and that the model successfully captures the dynamics of the time series. In the case of AdaBoost, strong and slowly decaying positive autocorrelations are observed across several lags before the break, suggesting that the model is unable to fully account for temporal dependencies. Although this pattern weakens after the break, short-term autocorrelation remains detectable. DLinear exhibits relatively low autocorrelation in both periods, with most values falling within the confidence bands, indicating a stable residual structure. The GRU and LSTM deep learning models display a similar pattern. Most autocorrelation coefficients are insignificant, suggesting that these methods effectively capture temporal dependencies both before and after the structural break. For Random Forest (RFR), pronounced positive autocorrelation emerges at lower lags following the break, indicating difficulties in adapting to structural changes. SVR exhibits more persistent positive autocorrelation across several lags after the break, implying that forecasting errors remain temporally dependent. TCN demonstrates relatively stable behavior, although moderate autocorrelation is observed at certain lags, particularly in the post-break period. For TFT, mild short-term autocorrelation appears before the break but diminishes afterward, with most lag values remaining within the confidence intervals.

The Taylor diagrams evaluate (Figure 3) model performance across three dimensions: correlation, standard deviation, and RMSE relative to the observed series. During the pre-break period of 2022, most models are located relatively close to the reference point, indicating high correlations and standard deviations similar to those of the actual time series. Deep learning methods, particularly GRU and LSTM, occupy especially favorable positions on the diagram, reflecting strong correlations and relatively low forecasting errors. ARIMA also demonstrates stable performance, although its standard deviation differs slightly from the reference value. Following the structural break, the positions of several methods move farther away from the ideal reference point, indicating performance deterioration. This effect is particularly evident among ensemble-based models, which exhibit declining correlations and increasing errors. AdaBoost and, to a lesser extent, Random Forest show greater deviations from the reference standard deviation, suggesting that they adapt less effectively to changing market dynamics. In contrast, GRU and LSTM remain relatively close to the reference point, indicating more robust forecasting capabilities even after the break. TCN and TFT exhibit intermediate performance: their correlations remain high, although both standard deviation and RMSE increase slightly.

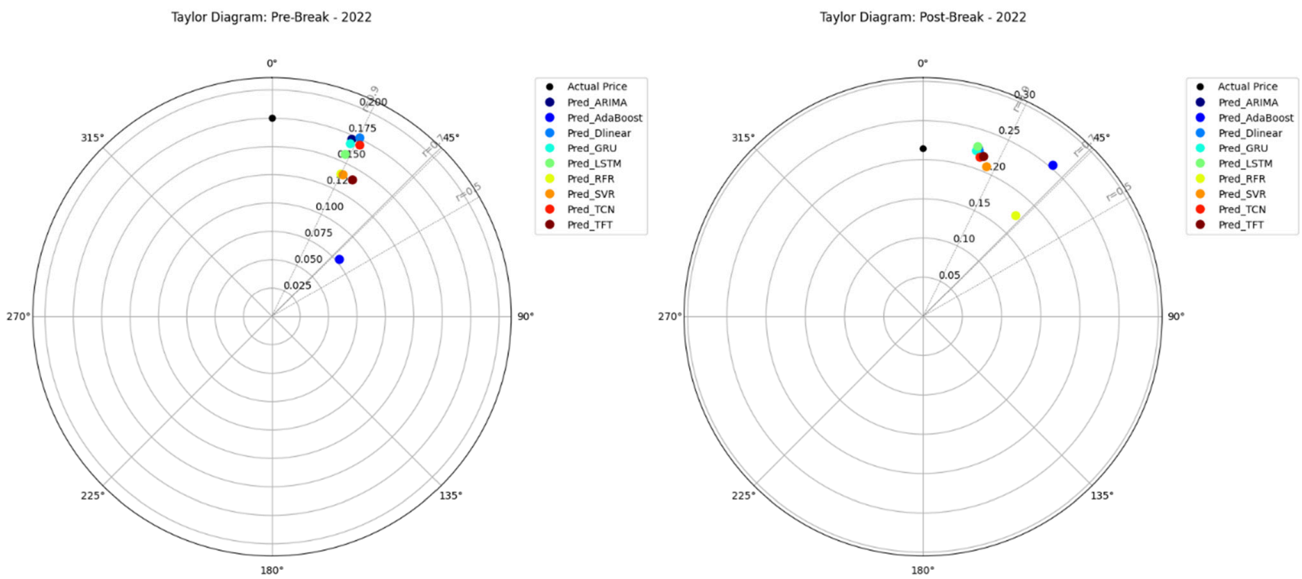

Figure 3. Taylor diagrams pre- and post-break in period 1 (2022). Source: authors' research.

The time-lagged cross-correlation (TLCC) analysis examines the correlation between model forecasts and the observed time series across different temporal lags (Figure 4). During the post-break period of 2022, the highest correlations for most methods occur at lags close to zero or slightly negative values. This suggests that model forecasts closely track the evolution of the actual time series, although minor temporal shifts are evident in some cases. Deep learning architectures, particularly GRU and LSTM, exhibit the highest correlation values, approaching 1.0 around lag −1. ARIMA and DLinear also demonstrate high correlations, albeit slightly lower than those of the deep learning models. In contrast, AdaBoost and SVR show a more rapid decline in correlation as positive lag values increase, suggesting a reduced ability to maintain a close relationship with the observed series over longer periods. Random Forest displays a similar pattern, although its correlations remain somewhat more stable. TCN and TFT exhibit intermediate performance, characterized by high correlations at lags near zero followed by a gradual decline at larger lag values.

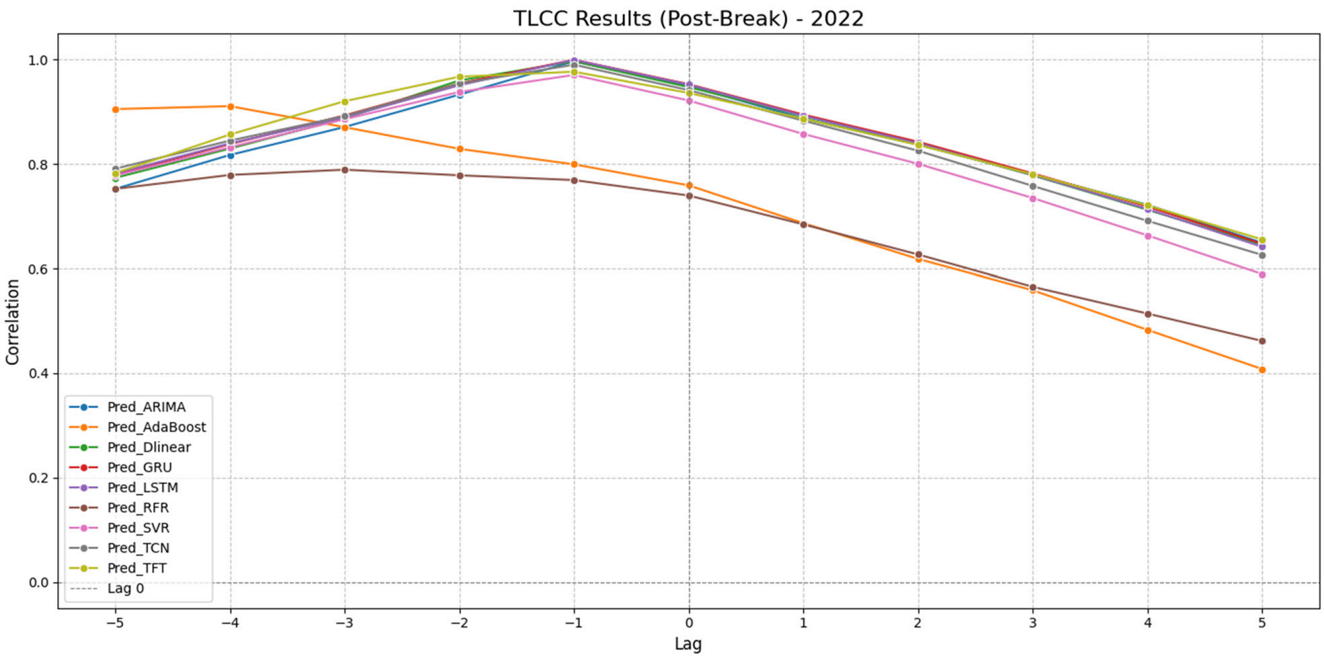

Figure 4. TLCC results post-break in period 1 (2022). Source: authors' research.

4.3.2. Structural Break Analysis for Period 2 (2025)

The rolling MAPE heatmap (Figure 5) also enables an assessment of the temporal evolution of forecasting errors during 2025. According to the figure, most models exhibit low forecasting errors during the first months of the year, as indicated by the predominance of green shading. During this period, models operated within a relatively stable market environment and were able to capture the patterns of the time series effectively. However, around the structural break marked by the red dashed line, several methods exhibit a sudden increase in forecasting errors, indicating that the structural change had a substantial impact on predictive performance. This effect is particularly pronounced for AdaBoost and Random Forest, where strong error spikes emerge around the break point, reflected by yellow and red color intensities. These findings suggest that ensemble-based models may be more sensitive to abrupt market changes and may require more time to adapt to new market dynamics. For DLinear, only moderate changes in forecasting errors are observed. Although a slight increase appears around the structural break, errors quickly return to lower levels. Deep learning architectures, especially GRU and LSTM, maintain relatively stable performance throughout the entire period, with errors remaining largely low and only moderate increases observed around the break point. SVR displays gradually increasing forecasting errors during certain segments of the post-break period, suggesting greater sensitivity to structural changes in the time series. TCN errors increase moderately near the break point but stabilize relatively quickly thereafter. TFT also exhibits a mild error spike around the structural break, followed by a return to lower error levels. Toward the end of the year, several models experience a renewed increase in forecasting errors, particularly the ensemble methods, suggesting that further changes in the market environment again challenged forecasting performance.

The autocorrelation function (ACF) plots (Appendix A.2) illustrate the temporal dependence of forecasting errors before and after the structural break using the 2025 dataset. For ARIMA, most autocorrelation coefficients remain within the confidence intervals in both periods, indicating residuals that closely resemble white noise and confirming the model's ability to capture time-series dynamics effectively. For AdaBoost, positive and slowly decaying autocorrelations are observed across several lags before the structural break, suggesting that the model does not fully account for temporal dependencies. Although this effect weakens after the break, some short-term autocorrelation remains detectable. DLinear exhibits low and rapidly decaying autocorrelations in both periods, with most values remaining within the confidence bands, indicating a stable residual structure. GRU and LSTM display a similar pattern, with most autocorrelation coefficients remaining statistically insignificant. These models effectively capture temporal dependencies both before and after the structural break. For Random Forest (RFR), mild positive autocorrelation emerges at lower lags after the break. SVR exhibits more persistent positive autocorrelation across multiple lags during the post-break period. TCN demonstrates relatively stable behavior, although moderate autocorrelation is present at some short lags, particularly after the break. For TFT, mild short-term autocorrelation is observed before the break but largely disappears afterward, with most values remaining within the confidence intervals.

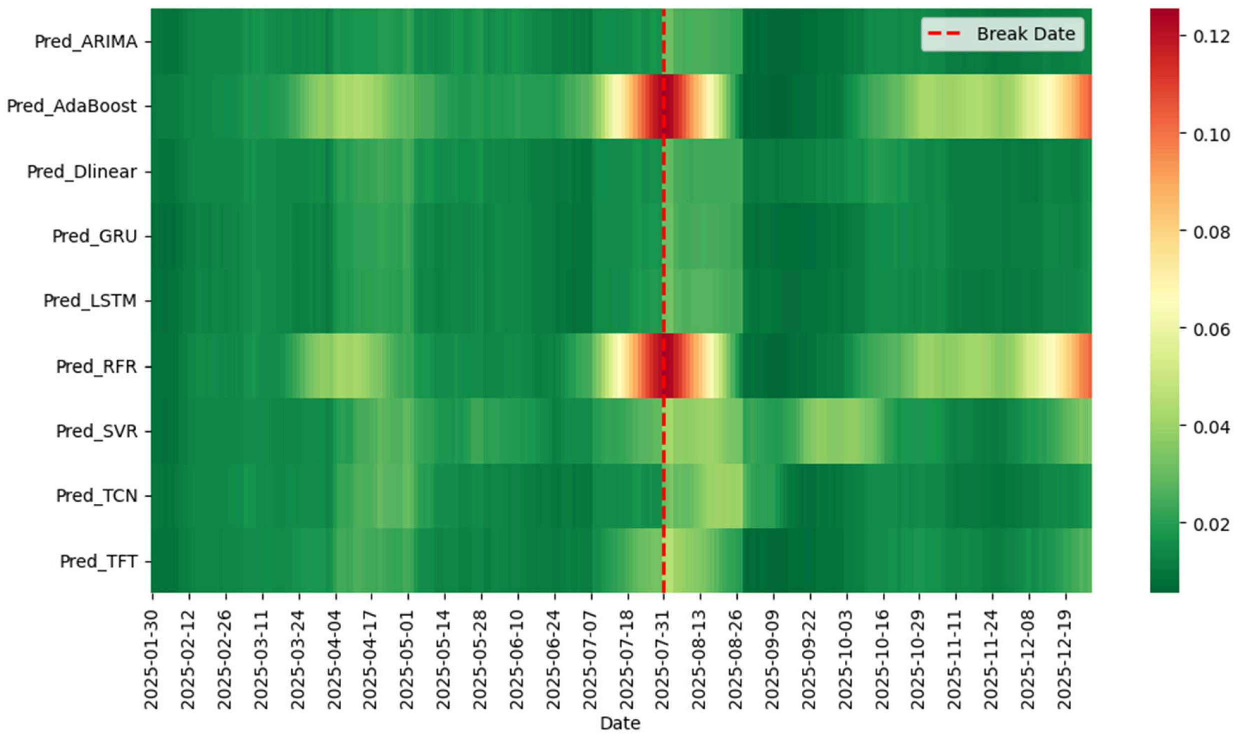

Figure 5. Model-level errors over time (2025) with structural break (5-day rolling MAPE). Source: authors' research.

The 2025 Taylor diagrams (Figure 6) reveal a performance structure similar to that observed in 2022, although differences among models become more pronounced. During the pre-break period, most methods exhibit high correlations with the observed time series and are located relatively close to the reference point. GRU, LSTM, and to some extent TFT occupy particularly favorable positions, indicating their ability to capture both the variability and dynamics of the time series effectively. ARIMA and DLinear also demonstrate stable performance, although their standard deviations differ slightly from the reference value in some cases. Following the structural break, performance differences among models become more distinct. Ensemble-based approaches, particularly AdaBoost and Random Forest, move farther away from the reference point, indicating increased forecasting errors. In contrast, deep learning models—especially GRU and LSTM—continue to exhibit high correlations and relatively low RMSE values. TCN and TFT occupy intermediate positions, displaying moderate deviations from the reference standard deviation. The diagrams also reveal that the impact of the structural break is reflected in model variance, visible through the radial displacement of points.

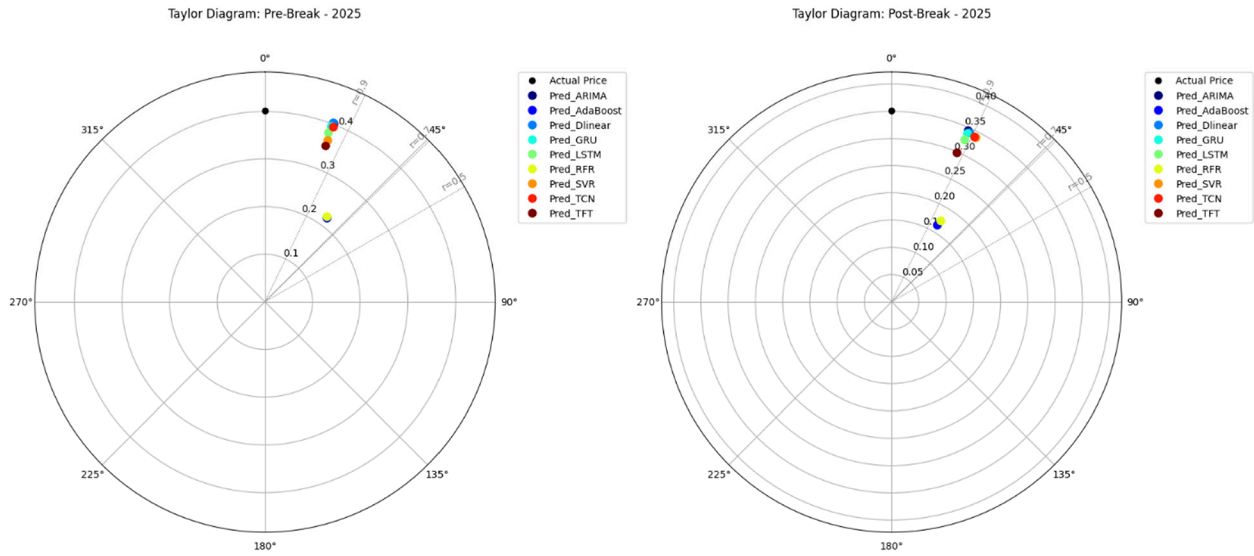

Figure 6. Taylor diagrams pre- and post-break in period 2 (2025). Source: authors' research.

The post-break TLCC diagram for 2025 (Figure 7) indicates that most models exhibit extremely high correlations at lag values close to zero, suggesting that forecasts closely follow the movements of the actual time series. In many cases, correlation coefficients approach 1.0, indicating a very strong predictive relationship. GRU, LSTM, and DLinear demonstrate particularly stable performance, as their correlations decline only slightly at larger positive or negative lag values. ARIMA and TCN also exhibit high correlations, although a gradual decline becomes apparent as lag increases. AdaBoost displays somewhat lower correlations, especially at larger positive lag values, indicating that its forecasts diverge more rapidly from the actual time series over time. Random Forest and SVR exhibit moderate stability. TFT demonstrates a relatively balanced pattern, characterized by high correlations near zero lag and moderate declines as lag increases.

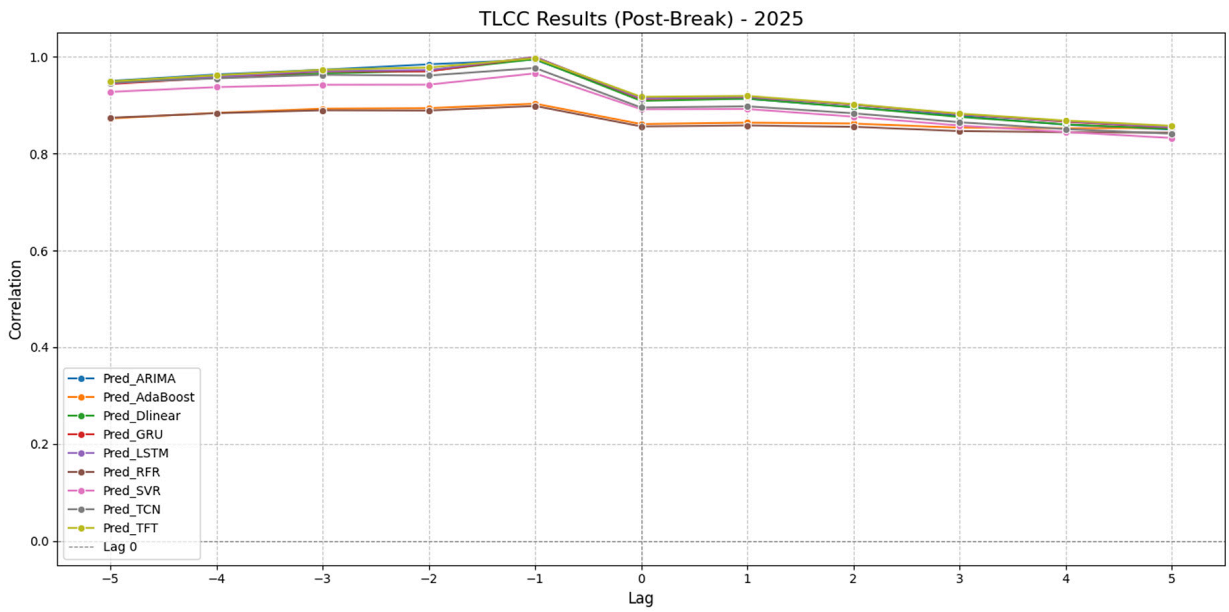

Figure 7. TLCC results post-break in period 2 (2025). Source: authors' research.

4.4. Analysis of Trading Strategies

4.4.1. Trading Strategy Results for Period 1 (2022)

The comparison of the performance of trading strategies applied during the 2022 period (Table 10) reveals substantial differences across both the six-month and twelve-month forecasting horizons. The buy-and-hold strategy generated negative cumulative returns in both cases (−17.27% over six months and −14.84% over twelve months). In contrast, several model-based strategies were able to generate positive returns, supporting the potential practical applicability of predictive models. For the six-month horizon, the best performance was achieved by the TFT-based strategy, which generated a cumulative return of 24.84% and a Sharpe ratio of 2.18, while maintaining a relatively low maximum drawdown (6.46%). The SVR model also delivered outstanding results, achieving a return of 19.75% and a Sharpe ratio of 1.76. AdaBoost and LSTM likewise produced positive returns, although their performance was somewhat lower. By contrast, certain deep learning models, such as GRU, generated negative cumulative returns and negative Sharpe ratios, suggesting that not all complex architectures are capable of delivering stable performance under the examined market conditions. At the twelve-month forecasting horizon, the differences among models became even more pronounced. The highest cumulative returns were achieved by strategies based on SVR and Random Forest (RFR) forecasts, generating returns of 35.72% and 36.57%, respectively, while also exhibiting exceptionally high Sharpe ratios. TFT also demonstrated strong performance, with a cumulative return of 32.95% and a Sharpe ratio of 1.50. In contrast, several methodologies—including ARIMA, GRU, and TCN—produced negative returns and were therefore less capable of adapting to the market conditions observed in 2022. Risk metrics further indicate that the best-performing models generally achieved relatively lower maximum drawdowns, suggesting a more favorable risk–return profile. The win rates reinforce this conclusion, as the top-performing models typically achieved values between 53% and 58%, indicating stable trading performance. Overall, the results suggest that advanced machine learning models (SVR, RFR, and TFT) may be capable of generating substantial excess returns relative to a passive buy-and-hold strategy during the examined period, even when their absolute forecasting accuracy is not among the most competitive.

Table 10. Trading strategy results for Period 1 (2022)—6- and 12-month horizon.

	6 Months				12 Months			
	Cumulative Return	Sharpe Ratio	Max Draw-down	Win Rate	Cumulative Return	Sharpe Ratio	Max Draw-down	Win Rate
Buy_and_Hold	−17.2651	−1.3383	28.2892	46.3415	−14.8447	−0.5300	42.8709	49.2000
ARIMA	4.4450	0.3946	9.0806	49.5935	−12.6254	−0.5731	28.2488	48.0000
AdaBoost	11.8752	1.0562	12.5344	55.2846	8.1161	0.3683	16.5917	52.0000
Dlinear	1.1587	0.1199	12.2181	52.0325	6.2893	0.2854	14.9626	50.0000
GRU	−6.7237	−0.5971	15.5297	49.5935	−6.8991	−0.3130	22.5136	49.6000
LSTM	8.0603	0.8352	8.6378	52.0325	4.0542	0.1839	15.8360	49.2000
RFR	7.1290	0.6332	14.1120	54.4715	36.5697	1.6681	12.9650	52.4000
SVR	19.7510	1.7637	12.5344	56.0976	35.7239	1.8284	11.0851	54.8000
TCN	2.2196	0.2068	16.4644	53.6585	−8.6443	−0.3712	24.0577	49.2000
TFT	24.8377	2.1803	6.4623	57.7236	32.9533	1.5016	11.7422	53.6000

Source: authors' research.

4.4.2. Trading Strategy Results for Period 2 (2025)

The results of trading strategies applied in the 2025 period (Table 11) differ substantially from the performance patterns observed in 2022. This indicates that changes in the market environment significantly influenced the effectiveness of the various models. Over the six-month forecasting horizon, the buy-and-hold strategy performed particularly well, achieving a cumulative return of 23.20% and a Sharpe ratio of 1.60. During this period, the copper market exhibited a stronger upward trend, which was favorable for passive investment strategies. However, this environment proved more challenging for several model-based approaches. Within the six-month horizon, the best-performing methods were LSTM- and GRU-based strategies. The LSTM model achieved a cumulative return of 18.94% and a Sharpe ratio of 1.74, while the GRU model delivered an 11.50% return and a Sharpe ratio of 1.05, with relatively moderate maximum drawdown. These results suggest that deep learning models were more effective at capturing market trends during this period. In contrast, several other models (such as AdaBoost, SVR, TCN, and TFT) produced negative cumulative returns and negative Sharpe ratios, indicating difficulties in capturing short-term market dynamics. Over the twelve-month forecasting horizon, even more pronounced differences emerged among model-based strategies. The passive buy-and-hold strategy itself achieved an outstanding return of 36.18%, but both LSTM and GRU were able to outperform it. The best performance was obtained by the LSTM model, which generated a cumulative return of 64.98% and a Sharpe ratio of 2.59, indicating an exceptionally strong risk–return trade-off. The GRU model also performed strongly, achieving a 49.28% cumulative return and a Sharpe ratio of 2.22. The DLinear model also delivered relatively stable performance, with a cumulative return of 31.22% and a Sharpe ratio of 1.26, suggesting that even simpler architectures can provide effective predictions under favorable market conditions. In contrast, several models performed significantly worse over the longer horizon. AdaBoost generated substantial losses across both time horizons, while SVR produced near-zero cumulative returns. The Random Forest–based strategy also showed weak performance, particularly over the longer forecasting period. Risk metrics indicate that the best-performing strategies generally maintained relatively low maximum drawdowns, suggesting a favorable risk–return profile. Win rates further support this observation: the strongest models (particularly LSTM and GRU) achieved values close to or above 58–59%. Overall, the results suggest that in the 2025 period, deep learning models (especially LSTM and GRU architectures) achieved the highest trading performance, while several traditional machine learning models significantly underperformed.

Table 11. Trading strategy results for Period 2 (2025)—6- and 12-month horizon.

	6 Months				**12 Months**			
	Cumulative Return	**Sharpe Ratio**	**Max Drawdown**	**Win Rate**	**Cumulative Return**	**Sharpe Ratio**	**Max Drawdown**	**Win Rate**
Buy_and_Hold	23.2005	1.5954	23.4059	59.8361	36.1864	0.9230	29.1312	56.4000
ARIMA	−2.1951	−0.2008	17.4684	51.6393	11.0513	0.3999	20.8990	52.0000
AdaBoost	−25.3530	−1.9817	32.1772	43.4426	−23.2685	−0.9273	34.7142	48.4000
Dlinear	−1.8923	−0.1591	24.4749	51.6393	31.2220	1.2577	26.6133	56.4000
GRU	11.5017	1.0543	14.3383	55.7377	49.2829	2.2228	15.8234	59.2000
LSTM	18.9413	1.7429	14.4366	55.7377	64.9815	2.5917	19.8173	58.4000
RFR	2.6820	0.2080	26.7889	50.0000	−8.8474	−0.3521	26.7889	48.8000

Table 11. *Cont.*

	6 Months				12 Months			
	Cumulative Return	Sharpe Ratio	Max Drawdown	Win Rate	Cumulative Return	Sharpe Ratio	Max Drawdown	Win Rate
SVR	−17.1721	−1.4519	21.8821	45.9016	−0.0076	−0.0003	24.0259	50.0000
TCN	−11.1165	−0.9376	24.3142	50.8197	7.5504	0.3231	20.0229	55.6000
TFT	−12.4188	−1.0335	26.2411	47.5410	2.2844	0.0909	26.2411	50.0000

Source: authors' research.

4.5. Statistical Validation of the Error–Profit Paradox via Rank Correlation

The results in Table 12 provide direct statistical evidence that superior forecast accuracy does not guarantee superior trading performance. In the 2022 sample, GRU achieved the lowest MAPE among all nine models (1.30%), yet its trading strategy produced a negative cumulative return in both the 6-month (−6.72%) and 12-month (−6.90%) windows. By contrast, SVR and TFT—whose forecast accuracy ranked sixth and seventh, respectively—generated cumulative returns of 19.75% and 24.84% (6-month) and 35.72% and 32.95% (12-month). This inverse pattern is not incidental: the Spearman correlation between the MAPE ranking and both trading-performance rankings is positive and statistically significant at the 12-month horizon ($\rho = 0.78$, $p = 0.013$ for Sharpe ratio; $\rho = 0.82$, $p = 0.007$ for cumulative return), and positive though only marginally significant at the 6-month horizon ($\rho = 0.65$, $p = 0.058$ for both measures). Taken together, these results confirm, with formal statistical support rather than descriptive observation alone, that at least one of the two sample periods examined in this study exhibits a genuine Error–Profit Paradox: the model that minimizes forecast error is not the model that maximizes economic return. The 2025 sample, where the correlation reverses sign and becomes significantly negative, further demonstrates that this disconnect is regime-dependent rather than a fixed property of any individual model.

Table 12. Spearman rank-correlation results 2022 and 2025.

MAPE vs. Sharpe Ratio				
Period	**Horizon**	**Spearman ρ**	***p*-Value**	**Interpretation**
2022	6 months	0.6500	0.058	Positive, marginally significant → paradox present
2022	12 months	0.7833	0.013	Positive, significant → paradox confirmed
2025	6 months	−0.7167	0.030	Negative, significant → paradox absent
2025	12 months	−0.9667	<0.001	Negative, highly significant → accuracy and profitability aligned
MAPE vs. Cumulative return				
Period	**Horizon**	**Spearman ρ**	***p*-value**	**Interpretation**
2022	6 months	0.6500	0.058	Positive, marginally significant → paradox present
2022	12 months	0.8167	0.007	Positive, significant → paradox confirmed
2025	6 months	−0.7167	0.030	Negative, significant → paradox absent
2025	12 months	−0.9667	<0.001	Negative, highly significant → accuracy and profitability aligned

Source: authors' research.

5. Discussion

The combined application of the autocorrelation function (ACF), rolling MAPE heatmaps, Taylor diagrams, and time-lagged cross-correlation (TLCC) analyses provides a comprehensive assessment of forecasting performance and model robustness in the presence of structural changes observed in copper price dynamics. Results from the ACF analysis indicate that deep learning models, particularly GRU and LSTM architectures, generate residuals that most closely resemble white noise processes, suggesting a superior ability to capture temporal dependencies embedded in the series. This finding is consistent with prior evidence showing that recurrent neural networks, especially gated architectures, are effective in modeling long-term nonlinear dependencies that characterize financial and commodity market time series [34,35]. The rolling MAPE heatmaps further support this conclusion. GRU- and LSTM-based models maintain relatively stable error levels around the identified structural break, whereas ensemble methods such as AdaBoost and Random Forest exhibit more pronounced error spikes. This pattern suggests that although ensemble approaches may perform well under relatively stable conditions, they are more sensitive to regime shifts. Such evidence aligns with previous studies reporting deteriorating performance of tree-based models in non-stationary environments characterized by structural disruptions [31,32]. In contrast, deep learning architectures appear better equipped to adapt to evolving data-generating processes, a particularly important characteristic in commodity markets. Taylor diagram results lead to similar conclusions. Deep learning models are generally positioned closer to the ideal reference point, indicating stronger correlations with observed values, improved variance reproduction, and lower RMSE values. These findings reinforce the growing consensus that model evaluation should extend beyond conventional error metrics and also consider the ability to reproduce the dynamic characteristics of the underlying time series [78,79].

The present results suggest that GRU and LSTM models outperform competing approaches not only in predictive accuracy but also in preserving the statistical structure of copper price dynamics. Although forecasting performance deteriorates for most models following the structural break, GRU, LSTM, and to some extent TFT architectures continue to exhibit relatively stable predictive properties. This observation is particularly relevant because structural breaks are widely recognized as one of the primary sources of forecast degradation and reduced out-of-sample performance [69,70]. The relative robustness of deep learning models under such conditions suggests a stronger capacity to accommodate regime changes, a valuable feature in highly volatile commodity markets.

The TLCC analysis provides additional insights. For most models, peak correlations are observed at lags close to zero, indicating that forecasts are most effective in tracking short-term movements in the actual series. This result is consistent with the broader forecasting literature, which demonstrates that predictive power typically declines as the forecast horizon increases [83].

The interpretation of these multi-layered dimensions is directly tied to the Error–Profit Paradox, which constitutes the central contribution of this study. While an increasing body of literature argues that statistical accuracy and economic usefulness do not necessarily coincide [73,74], our integrated framework—culminating in the cross-layer correlation layer outlined in Figure 1—provides a formal mathematical confirmation of this phenomenon. By executing a formal Spearman's rank correlation test between the MAPE rankings and financial profitability metrics, we quantified this decoupling across distinct market states, yielding non-significant coefficients.

This statistically proven paradox is driven by a fundamental structural asymmetry: traditional metrics like MAPE penalize errors symmetrically, optimizing models toward historical variance minimization, whereas financial returns are inherently non-linear and governed by directional synchronization at market boundaries. Consequently, models that maintain low average errors but fail to capture critical macroeconomic turning points experience severe capital erosion due to lagged position reversals. Conversely, architectures that exhibit higher baseline errors but successfully align with phase shifts (as verified by our TLCC analysis) capture high-yielding macro trends. This systemic decoupling confirms that statistical precision cannot serve as a direct proxy for economic utility in non-stationary regimes, expanding the conceptual boundaries established by recent financial engineering literature [27].

The analysis of the copper market is particularly relevant from an industrial and strategic perspective. Copper represents a critical input for electrical and electronic systems, power transmission infrastructure, electric mobility technologies, and renewable energy systems. Ongoing electrification trends—including the expansion of electric vehicle adoption, grid modernization, and the integration of renewable energy sources—are exerting substantial demand pressure on copper markets, potentially increasing both volatility and the likelihood of structural changes. This observation is consistent with studies emphasizing that strategic commodity markets are increasingly shaped by macroeconomic developments and technological transitions [51,52]. The findings indicate that deep learning architectures, particularly GRU and LSTM models, provide more robust forecasting performance during periods characterized by structural breaks and changing market regimes. At the same time, the results demonstrate that model evaluation cannot be reduced solely to statistical error measures. As demonstrated by our integrated evaluation matrix, future asset management and industrial hedging strategies must transition toward multi-dimensional validation schemes where predictive accuracy, dynamic consistency, and economic viability are evaluated collectively to prevent the premature rejection of models possessing high operational utility.

6. Conclusions

The objective of this study was to examine how the performance of statistical and artificial intelligence–based forecasting models can be evaluated in a complex industrial commodity market such as copper, particularly under economic conditions characterized by severe structural breaks. The empirical results consistently demonstrate that conventional statistical evaluation criteria—especially symmetric metrics such as Mean Absolute Percentage Error (MAPE)—do not necessarily reflect the economic usefulness of forecasting models. The empirical validation of this decoupling represents the central contribution of this study, confirming the presence of the Error–Profit Paradox. The underlying mechanism driving this paradox is rooted in the fundamental divergence between statistical loss functions and real-world economic utility. Traditional metrics like MAPE or RMSE penalize overestimations and underestimations symmetrically, optimizing models toward minimizing baseline variance and historical noise. However, trading profitability is inherently non-linear and asymmetrical, depending heavily on directional accuracy and the precise identification of market turning points rather than absolute numerical proximity. A model can exhibit a higher MAPE due to clustered, larger errors during prolonged, stable trend phases; yet, if it accurately predicts the critical inflection points and structural pivots where price velocity changes direction, the simulated trading strategy capitalizes on long or short positions at the most lucrative intervals. Conversely, a statistically precise model that minimizes average error but fails to capture these macroeconomic turning points will experience severe capital erosion due to delayed position reversals and rigid phase shifts. Thus, in liquid commodity futures, directional synchronization at market boundaries holds significantly higher economic value than absolute statistical error minimization. To synthesize the scientific core of this study, a clear distinction must be made between our methodological contribution and the substantive empirical findings. Methodological Contribution: this study introduces an integrated, four-layered evaluation framework that establishes a new benchmark for commodity market forecasting. By structurally connecting numerical accuracy (MAPE), statistical stability (Taylor diagrams), dynamic phase-shift screening (Time-Lagged Cross-Correlation), and asymmetric trading simulations, this framework prevents the premature rejection of models that possess high economic utility despite lower statistical precision. This holistic evaluation matrix allows institutional decision-makers to select predictive architectures based on multi-dimensional, integrated performance profiles rather than isolated error metrics. Substantive Empirical Findings: Our findings reveal that advanced recurrent architectures (GRU and LSTM) demonstrate superior structural resilience and faster parameter recalibration under severe macroeconomic shocks compared to ensemble tree-based models (Random Forest and AdaBoost). Gated neural mechanisms effectively isolate underlying structural signals from extreme noise, maintaining stable, dominant positions within the Model Confidence Set (MCS) boundaries even immediately following major Zivot-Andrews structural break points. These empirical insights expand upon recent commodity literature [94,95], shifting the academic discourse from pure statistical optimization toward directional and economic alignment in non-stationary regimes. From a practical perspective, the framework offers vital implications for commodity traders, investors, and industry decision-makers. It demonstrates that model selection must account for dynamic, real-world constraints—such as autocorrelation structures in residuals (ACF), where residuals closer to white noise translate into stronger long-term trading outcomes—especially in the copper market, where price movements are closely linked to industrial demand, electrification trends, and global business cycles.

Limitations and Future Research

While this study offers comprehensive insights into model evaluation under structural shifts, several limitations must be explicitly outlined to establish boundaries for generalizability and guide future academic inquiries. First, the empirical validation was restricted to a single primary commodity asset—the copper futures market. Although copper serves as an ideal global macroeconomic bellwether closely tied to the green energy transition, future research should extend this four-layered framework to other liquid base metals, energy derivatives, and strategic battery minerals (such as lithium, nickel, and cobalt) to verify the universality of the Error–Profit Paradox across diverse asset classes. Second, the predictive architectures relied strictly on endogenous price data. Incorporating exogenous macroeconomic variables, such as global industrial production indexes, central bank interest rate decisions, and geopolitical sentiment indicators, represents a prominent avenue for future research to enhance model adaptability and weight recalibration during regime shifts. Third, the trading simulation operated under the simplifying assumption of a frictionless market, excluding transaction costs, brokerage commissions, fiscal obligations, and market slippage. While this theoretical control baseline was conceptually vital to isolate the core mathematical mechanics of the paradox, real-world trading frictions introduce a high degree of heterogeneity. Execution costs would exert an asymmetrical impact across the model spectrum: architectures characterized by high-frequency position switching based on predictive noise would face severe capital erosion, whereas models generating fewer, macro-directional signals would demonstrate higher resilience. Consequently, omitting these frictions represents a highly conservative baseline for our core thesis; the inclusion of realistic market frictions would theoretically widen the gap between statistical precision and economic returns, further reinforcing the structural validity of the paradox. Fourth, our framework utilized a fixed 50-day rolling window specification and discrete 6- and 12-month forecasting horizons. Future extensions should conduct extensive sensitivity analyses across varying window lengths and short-to-medium investment horizons to fully stress-test the robustness of these deep learning architectures. Finally, future research could focus on the integrated optimization of forecasting models and trading strategies through the development of novel loss functions capable of simultaneously penalizing statistical variance and economic misalignments.

Author Contributions: Conceptualization, L.V., T.T. and T.B.; methodology, L.V.; software, L.V.; validation, L.V., T.T. and T.B.; formal analysis, L.V.; investigation, T.T. and T.B.; resources, L.V.; data curation, L.V.; writing—original draft preparation, L.V.; writing—review and editing, L.V., T.T. and T.B.; visualization, L.V.; supervision, T.T.; project administration, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data used in the present study are publicly available.

Acknowledgments: This research was carried out within the framework of the AI-Assisted Supply Chains and Marketing Research Group at the Hungarian University of Agriculture and Life Sciences (MATE). During the preparation of this manuscript, the authors used [Gemini 3.5 and ChatGPT-5.5] for the purposes of academic text proofreading, translation and language refinement, paraphrasing, academic writing guidance, and conceptual clarifications to improve clarity and coherence. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation Function
AdaBoost	AdaBoost
ADF	Augmented Dickey–Fuller Test
AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
COVID-19	Coronavirus Disease 2019
DFGLS	Dickey–Fuller Generalized Least Squares Test
DM	Diebold–Mariano Test
DLinear	Decomposition Linear Model
GRU	Gated Recurrent Unit
KPSS	Kwiatkowski–Phillips–Schmidt–Shin Test
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCS	Model Confidence Set
PP	Phillips–Perron Test
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SVR	Support Vector Regression
TCN	Temporal Convolutional Network
TFT	Temporal Fusion Transformer
TLCC	Time-Lagged Cross-Correlation
ZA	Zivot–Andrews Test

Appendix A

Appendix A.1

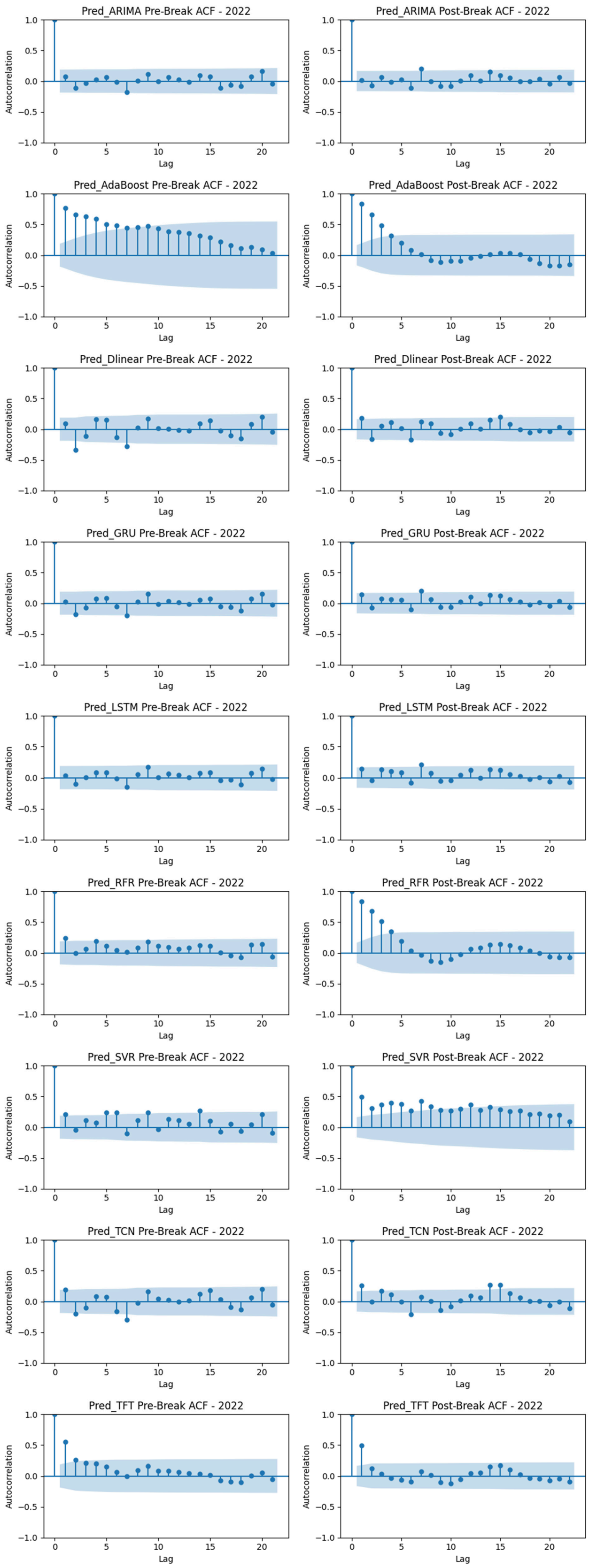

Figure A1. ACF Results Pre- and Post-Break in 2022 Source: Authors' Research.

https://doi.org/10.3390/make8070209

Appendix A.2

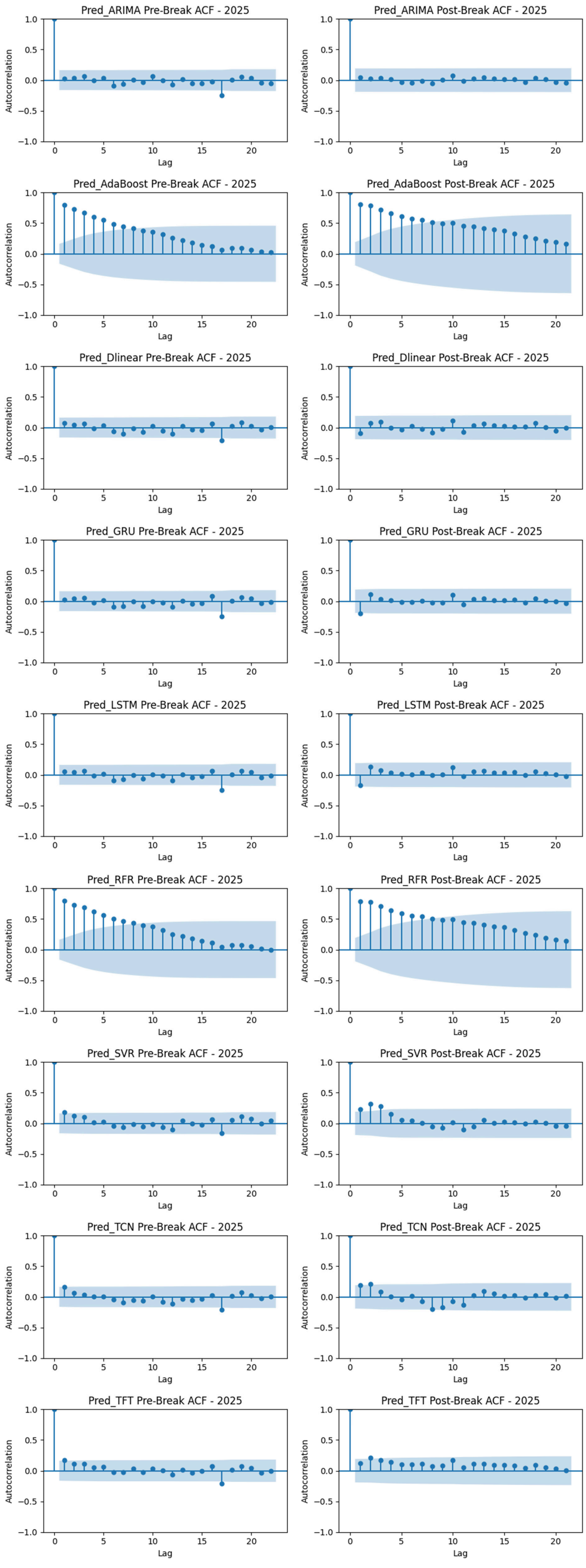

Figure A2. ACF Results Pre- and Post-Break in 2025 Source: Authors' Research.

References

1. Li, B.; Li, H.; Dong, Z.; Lu, Y.; Liu, N.; Hao, X. The global copper material trade network and risk evaluation: An industry chain perspective. *Resour. Policy* **2021**, *74*, 102275. [CrossRef]
2. Sugimura, Y.; Kawasaki, T.; Murakami, S. Potential for increased use of secondary raw materials in the copper industry as a countermeasure against climate change in Japan. *Sustain. Prod. Consum.* **2023**, *35*, 275–286. [CrossRef]
3. Podobińska-Staniec, M.; Wiktor-Sułkowska, A.; Kustra, A.; Lorenc-Szot, S. Copper as a critical resource in the energy transition. *Energies* **2025**, *18*, 969. [CrossRef]
4. Mahnoor, M.; Chandio, R.; Inam, A.; Ahad, I.U. Critical and strategic raw materials for energy storage devices. *Batteries* **2025**, *11*, 163. [CrossRef]
5. Rajaperumal, T.A.; Columbus, C.C. Transforming the electrical grid: The role of AI in advancing smart, sustainable, and secure energy systems. *Energy Inform.* **2025**, *8*, 51. [CrossRef]
6. Soares, A.F.; Spers, R.G.; Jhunior, R.D.O.S. Projection of global copper demand in the context of energy transition. *Resour. Policy* **2025**, *103*, 105567. [CrossRef]
7. Ushie, O.J.; Nwokolo, S.C.; Ohiero, P.O.; Iwuji, P.C.; Eyime, E.E.; Ogbulezie, J.C. Resilient strategies to mitigate the volatility of copper and aluminium critical mineral demand for net-zero electricity network production. *Sustain. Energy Res.* **2025**, *12*, 56. [CrossRef]
8. Basdekis, C.; Christopoulos, A.G.; Katsampoxakis, I.; Xanthopoulos, S. Trends and challenges after the impact of COVID-19 and the energy crisis on financial markets. *Energies* **2024**, *17*, 3857. [CrossRef]
9. Yang, S.; Fu, Y. Interconnectedness among supply chain disruptions, energy crisis, and oil market volatility on economic resilience. *Energy Econ.* **2025**, *143*, 108290. [CrossRef]
10. Aladwani, J. Influence of oil price fluctuations on inflation uncertainty. *J. Chin. Econ. Foreign Trade Stud.* **2025**, *18*, 44–85. [CrossRef]
11. Ben Ameur, H.; Boubaker, S.; Ftiti, Z.; Louhichi, W.; Tissaoui, K. Forecasting commodity prices: Empirical evidence using deep learning tools. *Ann. Oper. Res.* **2024**, *339*, 349–367. [CrossRef] [PubMed]
12. Rao, A.; Tedeschi, M.; Mohammed, K.S.; Shahzad, U. Role of economic policy uncertainty in energy commodities prices forecasting: Evidence from a hybrid deep learning approach. *Comput. Econ.* **2024**, *64*, 3295–3315. [CrossRef]
13. Manogna, R.L.; Dharmaji, V.; Sarang, S. Enhancing agricultural commodity price forecasting with deep learning. *Sci. Rep.* **2025**, *15*, 20903. [CrossRef] [PubMed]
14. Doroshenko, L.; Mastroeni, L.; Mazzoccoli, A. Wavelet and deep learning framework for predicting commodity prices under economic and financial uncertainty. *Mathematics* **2025**, *13*, 1346. [CrossRef]
15. Ao, X.; Gong, Y.; He, A. A review of time series prediction models based on deep learning. *IEEE Access* 2025, *in press*. [CrossRef]
16. Kong, X.; Chen, Z.; Liu, W.; Ning, K.; Zhang, L.; Muhammad Marier, S.; Xia, F. Deep learning for time series forecasting: A survey. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 5079–5112. [CrossRef]
17. Saravana, M.K.; Roopa, M.S.; Arunalatha, J.S.; Venugopal, K.R. Transformers for multivariate time series forecasting: Comprehensive analysis, challenges, research opportunities and future prospects. *IEEE Access* **2026**, *14*, 11424–11457. [CrossRef]
18. Ghoddusi, H.; Creamer, G.G.; Rafizadeh, N. Machine learning in energy economics and finance: A review. *Energy Econ.* **2019**, *81*, 709–727. [CrossRef]
19. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
20. Fischer, T.; Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **2018**, *270*, 654–669. [CrossRef]
21. Lim, B.; Arik, S.O.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
22. Ahmar, A.S.; Alfairus, M.Q.; Nursya'bani, N. Sustainable energy risk management: An integrated exponential smoothing and ARCH-GARCH framework for probabilistic forecasting. *Dev. Sustain. Econ. Financ.* **2025**, *8*, 100087. [CrossRef]
23. Al-Haddad, A.; Abu-Rayash, A. Volatility modeling in energy commodity markets: A review of responses to shocks and the path towards resiliency and sustainability. *Appl. Energy* **2026**, *411*, 127559. [CrossRef]
24. Maghyereh, A.; Ziadat, S.A. Dynamic interactions in futures markets: Exploring transitory and persistent intraday volatility linkages among oil, gold, stocks, and forex markets. *Comput. Econ.* **2026**, 1–47. [CrossRef]
25. Chen, J.; Yi, J.; Liu, K.; Cheng, J.; Feng, Y.; Fang, C. Copper price prediction using LSTM recurrent neural network integrated simulated annealing algorithm. *PLoS ONE* **2023**, *18*, e0285631. [CrossRef] [PubMed]
26. Derakhshani, R.; GhasemiNejad, A.; Amani Zarin, N.; Amani Zarin, M.M.; Jalaee, M.S. Forecasting copper prices using deep learning: Implications for energy sector economies. *Mathematics* **2024**, *12*, 2316. [CrossRef]
27. Zhang, C.; Sjarif, N.N.A.; Ibrahim, R. Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *WIREs Data Min. Knowl. Discov.* **2024**, *14*, e1519. [CrossRef]

28. Olubusola, O.; Mhlongo, N.Z.; Daraojimba, D.O.; Ajayi-Nifise, A.O.; Falaiye, T. Machine learning in financial forecasting: A US review: Exploring the advancements, challenges, and implications of AI-driven predictions in financial markets. *World J. Adv. Res. Rev.* **2024**, *21*, 1969–1984. [CrossRef]
29. Sahu, S.K.; Mokhade, A.; Bokde, N.D. An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Appl. Sci.* **2023**, *13*, 1956. [CrossRef]
30. Gao, H.; Kou, G.; Liang, H.; Zhang, H.; Chao, X.; Li, C.C.; Dong, Y. Machine learning in business and finance: A literature review and research opportunities. *Financ. Innov.* **2024**, *10*, 86. [CrossRef]
31. Han, H.; Liu, Z.; Barrios Barrios, M.; Li, J.; Zeng, Z.; Sarhan, N.; Awwad, E.M. Time series forecasting model for non-stationary series pattern extraction using deep learning and GARCH modeling. *J. Cloud Comput.* **2024**, *13*, 2. [CrossRef]
32. Osman, E.G.A.; Otaibi, F.A. Integrating deep learning and econometrics for stock price prediction: A comprehensive comparison of LSTM, transformers, and traditional time series models. *Mach. Learn. Appl.* **2025**, *22*, 100730. [CrossRef]
33. Dawi, N.B.M.; Matejicek, M.; Maresova, P. Unraveling financial market dynamics: The application of fractal theory in financial time series analysis. *Fractals* **2025**, *33*, 2550017. [CrossRef]
34. Uygun, Y.; Sefer, E. Financial asset price prediction with graph neural network-based temporal deep learning models. *Neural Comput. Appl.* **2025**, *37*, 25445–25471. [CrossRef]
35. Sanaei, B.; Daneshvar, S. Enhancing Forex market forecasting with ConvLSTM2D: A comprehensive analysis of spatiotemporal dependencies and data preprocessing techniques. *Comput. Econ.* **2025**, *67*, 3021–3065. [CrossRef]
36. Matejko, M.; Braciník, P. Comprehensive analysis of weather and commodity impacts on day-ahead electricity market using public API data with development of an accessible forecasting mode. *Electricity* **2026**, *7*, 10. [CrossRef]
37. Cahuantzi, R.; Chen, X.; Güttel, S. A comparison of LSTM and GRU networks for learning symbolic sequences. In *Science and Information Conference*; Springer Nature: Cham, Switzerland, 2023; pp. 771–785. [CrossRef]
38. Yunita, A.; Pratama, M.I.; Almuzakki, M.Z.; Ramadhan, H.; Akhir, E.A.P.; Mansur, A.B.F.; Basori, A.H. Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models. *MethodsX* **2025**, *15*, 103462. [CrossRef] [PubMed]
39. Farhadi, A.; Zamanifar, A.; Alipour, A.; Taheri, A.; Asadolahi, M. A hybrid LSTM-GRU model for stock price prediction. *IEEE Access* 2025, *in press*. [CrossRef]
40. El-Meehy, A.O.; El-Kharbotly, A.K.; El-Beheiry, M.M. Systematic hyperparameter analysis of GRU and LSTM across demand pattern types: A demand-characteristic-driven meta-learning framework for rapid optimization. *Sci. Rep.* **2025**, *15*, 44657. [CrossRef] [PubMed]
41. Dai, W.; An, Y.; Long, W. Price change prediction of ultra high frequency financial data based on temporal convolutional network. *Procedia Comput. Sci.* **2022**, *199*, 1177–1183. [CrossRef]
42. Yu, Q.; Yang, G.; Wang, X.; Shi, Y.; Feng, Y.; Liu, A. A review of time series forecasting and spatio-temporal series forecasting in deep learning. *J. Supercomput.* **2025**, *81*, 1160. [CrossRef]
43. Bao, W.; Cao, Y.; Yang, Y.; Wen, S. Long short-term financial time series forecasting based on residual multiscale TCN sparse expert network and informer. *IEEE Trans. Neural Netw. Learn. Syst.* 2025, *in press*. [CrossRef] [PubMed]
44. Liu, Y.; Huang, X.; Xiong, L.; Chang, R.; Wang, W.; Chen, L. Stock price prediction with attentive temporal convolution-based generative adversarial network. *Array* **2025**, *25*, 100374. [CrossRef]
45. Liu, X.; Wang, W. Deep time series forecasting models: A comprehensive survey. *Mathematics* **2024**, *12*, 1504. [CrossRef]
46. Caetano, R.; Oliveira, J.M.; Ramos, P. Transformer-based models for probabilistic time series forecasting with explanatory variables. *Mathematics* **2025**, *13*, 814. [CrossRef]
47. Kim, J.; Kim, H.; Kim, H.; Lee, D.; Yoon, S. A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges. *Artif. Intell. Rev.* **2025**, *58*, 216. [CrossRef]
48. Lee, M.C. Temporal fusion transformer-based trading strategy for multi-crypto assets using on-chain and technical indicators. *Systems* **2025**, *13*, 474. [CrossRef]
49. Omole, O.; Enke, D. Deep learning for Bitcoin price direction prediction: Models and trading strategies empirically compared. *Financ. Innov.* **2024**, *10*, 117. [CrossRef]
50. Bhuiyan, M.S.M.; Rafi, M.A.; Rodrigues, G.N.; Mir, M.N.H.; Ishraq, A.; Mridha, M.F.; Shin, J. Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies. *Array* **2025**, *26*, 100390. [CrossRef]
51. Feix, T.; Hache, E. Cumulative energy demand and global warming potential of metals and minerals production: Assessment, projections and mitigation options. *Resour. Policy* **2025**, *102*, 105516. [CrossRef]
52. Sverdrup, H.U.; van Allen, O.; Haraldsson, H.V. Modeling indium extraction, supply, price, use and recycling 1930–2200 using the WORLD7 model: Implication for the imaginaries of sustainable Europe 2050. *Nat. Resour. Res.* **2024**, *33*, 539–570. [CrossRef]
53. Le, T.V.; Dente, S.M.R.; Hashimoto, S. Contemporary and future secondary copper reserves of Southeast Asian countries. *Recycling* **2024**, *9*, 116. [CrossRef]

54. Andersen, E.V.; Shan, Y.; Bruckner, B.; Černý, M.; Hidiroglu, K.; Hubacek, K. The vulnerability of shifting towards a greener world: The impact of the EU's green transition on material demand. *Sustain. Horiz.* **2024**, *10*, 100087. [CrossRef]
55. U.S. Geological Survey. *Mineral Commodity Summaries 2025*; U.S. Geological Survey: Reston, VA, USA, 2025. [CrossRef]
56. Islam, M.S.; Islam, M.M.; Rehman, A.U.; Alam, M.F.; Tarique, M. Mineral production amidst the economy of uncertainty: Response of metallic and non-metallic minerals to geopolitical turmoil in Saudi Arabia. *Resour. Policy* **2024**, *90*, 104824. [CrossRef]
57. Ashena, M.; Laal Khezri, H.; Shahpari, G. Investigation into the dynamic relationships between global economic uncertainty and price volatilities of commodities, raw materials, and energy. *Appl. Econ. Anal.* **2024**, *32*, 23–40. [CrossRef]
58. Khurshid, A.; Chen, Y.; Rauf, A.; Khan, K. Critical metals in uncertainty: How Russia–Ukraine conflict drives their prices? *Resour. Policy* **2023**, *85*, 104000. [CrossRef]
59. Ben-Salha, O.; Zmami, M.; Waked, S.S.; Najjar, F.; Alenazi, Y.M. On the time-varying spillover between nonferrous metals prices, geopolitical risks, and global economic policy uncertainty. *Econ. Change Restruct.* **2025**, *58*, 5. [CrossRef]
60. Hu, C.; Liu, X.; Pan, B.; Sheng, H.; Zhong, M.; Zhu, X.; Wen, F. The impact of international price shocks on China's nonferrous metal companies: A case study of copper. *J. Clean. Prod.* **2017**, *168*, 254–262. [CrossRef]
61. Wang, C.; Zhang, X.; Wang, M.; Lim, M.K.; Ghadimi, P. Predictive analytics of the copper spot price by utilizing complex network and artificial neural network techniques. *Resour. Policy* **2019**, *63*, 101414. [CrossRef]
62. Saint Akadiri, S.; Ozkan, O. Energy, critical minerals, and precious metals: Navigating interconnectedness and portfolio strategies in investment risk management. *Resour. Policy* **2025**, *110*, 105747. [CrossRef]
63. Aue, A.; Horváth, L. Structural breaks in time series. *J. Time Ser. Anal.* **2013**, *34*, 1–16. [CrossRef]
64. Karavias, Y.; Narayan, P.K.; Westerlund, J. Structural breaks in interactive effects panels and the stock market reaction to COVID-19. *J. Bus. Econ. Stat.* **2023**, *41*, 653–666. [CrossRef]
65. Acar, S.; Altıntaş, N.; Haziyev, V. The effect of financial development and economic growth on ecological footprint in Azerbaijan: An ARDL bound test approach with structural breaks. *Environ. Ecol. Stat.* **2023**, *30*, 41–59. [CrossRef]
66. Chen, B.; Hong, Y. Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica* **2012**, *80*, 1157–1183. [CrossRef]
67. Oyadeyi, O.O. The velocity of money and lessons for monetary policy in Nigeria: An application of the quantile ARDL approach. *J. Knowl. Econ.* **2025**, *16*, 8249–8285. [CrossRef]
68. Tabash, M.I.; Asad, M.; Khan, A.A.; Sheikh, U.A.; Babar, Z. Role of 2008 financial contagion in effecting the mediating role of stock market indices between the exchange rates and oil prices: Application of the unrestricted VAR. *Cogent Econ. Financ.* **2022**, *10*, 2139884. [CrossRef]
69. Benigno, G.; Foerster, A.; Otrok, C.; Rebucci, A. Estimating macroeconomic models of financial crises: An endogenous regime-switching approach. *Quant. Econ.* **2025**, *16*, 1–47. [CrossRef]
70. Chung, V.; Espinoza, J.; Quispe, R. Forecasting financial volatility under structural breaks: A comparative study of GARCH models and deep learning techniques. *J. Risk Financ. Manag.* **2025**, *18*, 494. [CrossRef]
71. Pakštaitė, V.; Filatovas, E.; Juodis, M.; Paulavičius, R. Bitcoin price regime shifts: A Bayesian MCMC and hidden Markov model analysis of macroeconomic influence. *Mathematics* **2025**, *13*, 1577. [CrossRef]
72. Livieris, I.E.; Pintelas, P. A novel multi-step forecasting strategy for enhancing deep learning models' performance. *Neural Comput. Appl.* **2022**, *34*, 19453–19470. [CrossRef]
73. Han, Y.; Kim, J.; Enke, D. A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost. *Expert Syst. Appl.* **2023**, *211*, 118581. [CrossRef]
74. Hossain, M.N.; Mita, T.A.B. An empirical study of big data-enabled predictive analytics and their impact on financial forecasting and market decision-making. *Rev. Appl. Sci. Technol.* **2024**, *3*, 143–182. [CrossRef]
75. Rahman, M.M.; Hossain, M.A. Impact of big data and predictive analytics on financial forecasting accuracy and decision-making in global capital markets. *Am. J. Sch. Res. Innov.* **2024**, *3*, 99–140. [CrossRef]
76. Ding, H.; Li, Y.; Wang, J.; Chen, H.; Guo, D.; Zhang, Y. Large language model agent in financial trading: A survey. *arXiv* **2024**, arXiv:2408.06361. [CrossRef]
77. Bai, Y.; Gao, Y.; Wan, R.; Zhang, S.; Song, R. A review of reinforcement learning in financial applications. *Annu. Rev. Stat. Appl.* **2025**, *12*, 209–232. [CrossRef]
78. Lee, M.C.; Chang, J.W.; Yeh, S.C.; Chia, T.L.; Liao, J.S.; Chen, X.M. Applying attention-based BiLSTM and technical indicators in the design and performance analysis of stock trading strategies. *Neural Comput. Appl.* **2022**, *34*, 13267–13279. [CrossRef] [PubMed]
79. Saud, A.S.; Shakya, S. Technical indicator empowered intelligent strategies to predict stock trading signals. *J. Open Innov. Technol. Mark. Complex.* **2024**, *10*, 100398. [CrossRef]
80. Sukma, N.; Namahoot, C.S. Enhancing trading strategies: A multi-indicator analysis for profitable algorithmic trading. *Comput. Econ.* **2025**, *65*, 3807–3840. [CrossRef]
81. Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]

82. Du, P.; Gui, F.; Shan, L.; Xing, Q.; Wang, J. Probabilistic forecasting of non-ferrous metal prices based on outlier treatment algorithms, quantile regression based deep learning and two-phase multi-objective optimization. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112277. [CrossRef]
83. Kong, P.; Sun, B.; Yang, H.; Huang, X. Nonferrous metal price forecasting based on signal decomposition and ensemble learning. *J. Process Control* **2024**, *133*, 103146. [CrossRef]
84. Kontopoulou, V.I.; Panagopoulos, A.D.; Kakkos, I.; Matsopoulos, G.K. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet* **2023**, *15*, 255. [CrossRef]
85. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
86. Ismail, M.S.; Noorani, M.S.M.; Ismail, M.; Razak, F.A.; Alias, M.A. Predicting next day direction of stock price movement using machine learning methods with persistent homology: Evidence from Kuala Lumpur Stock Exchange. *Appl. Soft Comput.* **2020**, *93*, 106422. [CrossRef]
87. Park, H.J.; Kim, Y.; Kim, H.Y. Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Appl. Soft Comput.* **2022**, *114*, 108106. [CrossRef]
88. Nikou, M.; Mansourfar, G.; Bagherzadeh, J. Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intell. Syst. Account. Financ. Manag.* **2019**, *26*, 164–174. [CrossRef]
89. Nabipour, M.; Nayyeri, P.; Jabani, H.; Shahab, S.; Mosavi, A. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data: A comparative analysis. *IEEE Access* **2020**, *8*, 150199–150212. [CrossRef]
90. Liu, J.; Gong, C.; Chen, S.; Zhou, N. Multi-step-ahead wind speed forecast method based on outlier correction, optimized decomposition, and DLinear model. *Mathematics* **2023**, *11*, 2746. [CrossRef]
91. Dorbane, A.; Harrou, F.; Sun, Y. Exploring deep learning methods to forecast mechanical behavior of FSW aluminum sheets. *J. Mater. Eng. Perform.* **2023**, *32*, 4047–4063. [CrossRef]
92. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
93. Zheng, R.; Bao, Y.; Zhao, L.; Xing, L. Method to predict alloy yield based on multiple raw material conditions and a PSO-LSTM network. *J. Mater. Res. Technol.* **2023**, *27*, 3310–3322. [CrossRef]
94. Zhang, J.; Li, Z.; Wang, E.; Yu, B.; Li, J.; Ma, J. A temporal network based on characterizing and extracting time series in copper smelting for predicting matte grade. *Sensors* **2024**, *24*, 7492. [CrossRef] [PubMed]
95. Arrar, D.; Ounoughi, C.; Kamel, N.; Kalvet, T.; Tiits, M.; Ben Yahia, S. High-accuracy prediction of international raw material trade flows using temporal fusion transformer. *Int. J. Data Sci. Anal.* **2025**, *20*, 5631–5651. [CrossRef]
96. Becerra, M.; Jerez, A.; Garces, H.O.; Demarco, R. Copper price: A brief analysis of China's impact over its short-term forecasting. *Resour. Policy* **2022**, *75*, 102449. [CrossRef]
97. Oikonomou, K.; Damigos, D.; Dimitriou, D. Globality in the metal markets: Leveraging cross-learning to forecast aluminum and copper prices. *Resour. Policy* **2025**, *103*, 105558. [CrossRef]